|  |
| --- |
| BAN CƠ YẾU CHÍNH PHỦ  **HỌC VIỆN KỸ THUẬT MẬT MÃ**    ĐỒ ÁN TỐT NGHIỆP  **XÂY DỰNG VÀ ĐẢM BẢO AN TOÀN CHO HỆ THỐNG THU THẬP VÀ CHIA SẺ TIN TỨC TRỰC TUYẾN**  Ngành: An toàn thông tin  Mã số: 7.48.02.02  *Sinh viên thực hiện*:  **Phạm Hảo Nam**  Lớp: AT13G  *Người hướng dẫn*:  **TS. Trần Anh Tú**  Khoa An toàn thông tin – Học viện Kỹ thuật mật mã    **Hà Nội, tháng 2 năm 2021** |

[DANH MỤC HÌNH VẼ 6](#_Toc74985178)

[LỜI CẢM ƠN 7](#_Toc74985179)

[CHƯƠNG 1. GIỚI THIỆU CHUNG VỀ ĐỀ TÀI 8](#_Toc74985180)

[1. Thực trạng hiện nay 8](#_Toc74985181)

[2. Giải pháp 9](#_Toc74985182)

[3. Các vấn đề an toàn 9](#_Toc74985183)

[4. Kết luận 10](#_Toc74985184)

[CHƯƠNG 2: KHẢO SÁT VÀ PHÂN TÍCH HỆ THỐNG 11](#_Toc74985185)

[1.PHÂN TÍCH YÊU CẦU 11](#_Toc74985186)

[1.1. Chức năng 11](#_Toc74985187)

[1.2. Yêu cầu đặt ra 11](#_Toc74985188)

[2. PHÂN TÍCH HỆ THỐNG 12](#_Toc74985189)

[2.1 Biểu đồ tổng quát hệ thống 12](#_Toc74985190)

[2.2 Cơ sở dữ liệu 13](#_Toc74985191)

[2.3 Hệ thống thu thập thông tin 14](#_Toc74985192)

[2.4 Hệ thống hiển thị dữ liệu cho người dùng 16](#_Toc74985193)

[CHƯƠNG 3: THỰC HIỆN VÀ TRIỂN KHAI HỆ THỐNG WEB TIN TỨC 17](#_Toc74985194)

[1. Giao diện triển khai 17](#_Toc74985195)

[1.1 Giao diện trang chủ 17](#_Toc74985196)

[1.2 Giao diện bài báo chi tiết 18](#_Toc74985197)

[1.3 Giao diện admin 19](#_Toc74985198)

[CHƯƠNG 4. TRIỂN KHAI WEB CRAWLING 19](#_Toc74985199)

[1. Giới thiệu 19](#_Toc74985200)

[2. Đặc điểm website 19](#_Toc74985201)

[3. Web sẽ có sự thay đổi thường xuyên 20](#_Toc74985202)

[4. Lập chỉ mục và truy vấn các web 21](#_Toc74985203)

[5. Các lỗi về web crawling 22](#_Toc74985204)

[5.1 Chính sách tuyển chọn 23](#_Toc74985205)

[5.2 Chính sách về tái cập nhật trang web 23](#_Toc74985206)

[5.3 Chính sách chạy song song 23](#_Toc74985207)

[6. Kiến trúc web crawler 24](#_Toc74985208)

[6.1. Các vấn đề liên quan tới lập lịch 24](#_Toc74985209)

[6.2. Các vấn đề liên quan tới mô hình hoạt động 25](#_Toc74985210)

[6.3. Tách biệt ngắn hạn với lập kế hoạch dài hạn 27](#_Toc74985211)

[6.4. Kết hợp chất lượng và độ mới của trang 27](#_Toc74985212)

[7. Thu thập thông tin trên Web Vô hạn 29](#_Toc74985213)

[7.1 Trang web tĩnh và trang web động 29](#_Toc74985214)

[7.2 Mô hình người lướt ngẫu nhiên cho một trang web vô hạn 31](#_Toc74985215)

[7.3 Mô hình A: quay lại từng cấp một 33](#_Toc74985216)

[7.4 Dữ liệu từ các phiên người dùng trong các trang Web 34](#_Toc74985217)

[8. Đề xuất cho Hợp tác Máy chủ Web 35](#_Toc74985218)

[8.1 Đề án hợp tác giữa các máy chủ 35](#_Toc74985219)

[8.2 Hợp tác dựa trên lược đồ thăm dò 37](#_Toc74985220)

[8.3 Hợp tác dựa trên lược đồ ngắt 38](#_Toc74985221)

[9. Triển khai Crawler 39](#_Toc74985222)

[9.1. Môi trường lập trình và các phụ thuộc 39](#_Toc74985223)

[9.2 Về chương trình 39](#_Toc74985224)

[9.3 Vấn đề an toàn 41](#_Toc74985225)

[CHƯƠNG 5: TRIỂN KHAI CÁC BIÊN PHÁP AN TOÀN 43](#_Toc74985226)

[1. Tích hợp Load balancing 43](#_Toc74985227)

[2. Tích hợp tor proxy vào để thực hiện crawling 45](#_Toc74985228)

[3. Thực hiện việc phân quyền cho hệ thống 45](#_Toc74985229)

[4. Ngăn chặn các tấn công khác 46](#_Toc74985230)

[KẾT LUẬN 47](#_Toc74985231)

# DANH MỤC HÌNH VẼ

Hình 2.2.1: Sơ đồ thiết kế hệ thống

Hình 2.2.1: Cơ sở dữ liệu của Authen Service

Hình 2.2.3: Sơ đồ phân cấp chức năng hệ thống thu thập thông tin

Hình 2.2.4: Sơ đồ phân cấp chức năng hiển thị tin tức cho người dùng

Hình 3.1: Giao diện trang chủ người dùng

Hình 3.2: Giao diện bài báo chi tiết người dùng

Hình 3.3: Giao diện quản trị viên

Hình 3.4: Quy trình thực hiện việc thu thập dứ liệu

Hình 3.5: Cơ chế hoạt động của Load Balancing

Hình 3.6: Triển khai cấu hình routing cho website

Hình 3.7: Kết quả sử dụng proxy

Hình 3.8: Người dùng đăng nhập thất bại vào trang quản trị

# LỜI CẢM ƠN

Trong quá trình thực hiện đồ án tốt nghiệp này, tôi đã nhận được sự giúp đỡ tận tình của cán bộ hướng dẫn là TS. Trần Anh Tú – Giảng viên Khoa An toàn thông tin Học viện Kỹ thuật Mật mã, sự quan tâm sâu sát của cán bộ Hệ quản lý học viên quốc tế, sự động viên của người thân và bạn bè.

Xin cảm ơn tất cả mọi người đã tạo những điều kiện tốt nhất để tôi hoàn thành đồ án tốt nghiệp này!

|  |  |
| --- | --- |
|  | **SINH VIÊN THỰC HIỆN ĐỒ ÁN**  Phạm Hảo Nam |

# CHƯƠNG 1. GIỚI THIỆU CHUNG VỀ ĐỀ TÀI

## Thực trạng hiện nay

Bên cạnh những lợi ích mang lại, thì sự phổ biến của mạng xã hội ra đã tạo môi trường thuận lợi cho tin giả phát triển. Thực tế, trong những năm qua, không chỉ tại Việt Nam, mà trên hầu hết các quốc gia trên thế giới, tin giả đang hoành hành.

Tin giả (fake news) được hiểu là những thông tin không chính xác, chưa được kiểm chứng, kiểm duyệt, xuất hiện sai lệch về nội dung, được khuếch tán trên Internet và các phương tiện truyền thông. Một điều tệ hại là những thông tin sai lệch có xu hướng lan truyền nhanh hơn tin thật trên các dịch vụ truyền thông xã hội. Đây là một thực trạng đáng báo động, khiến các nhà chức trách, cơ quan thực thi pháp luật gặp nhiều khó khăn.

Trong khoảng vài năm trở lại đây, có hàng trăm tin giả xuất hiện trên báo chí ở Việt Nam, ảnh hưởng không nhỏ đến nhiều cá nhân, tổ chức và đời sống xã hội nói chung. Tin giả xuất hiện trên mọi lĩnh vực: Chính trị, kinh tế, văn hóa, xã hội… Mặc dù các cơ quan báo chí đều có quy trình kiểm chứng nguồn tin, trải qua nhiều bước kiểm duyệt, nhưng tin giả vẫn lọt qua hàng rào bảo vệ để hiện diện trên báo.

Dịch bệnh COVID-19 có lẽ là mảnh đất màu mỡ để tin giả bùng phát với quy mô chưa từng thấy trên thế giới và cả ở Việt Nam: từ những người vô tình hay hữu ý tung tin về việc phát hiện ca nhiễm bệnh ở chỗ này chỗ kia, từ những status vô cùng ngớ ngẩn nhưng vẫn thu hút khá đông lượt chia sẻ như chính phủ phun thuốc lên trời để chống dịch hoặc phải ăn 6-12 quả trứng luộc để phòng virus, cho đến những thông tin gây hoang mang và hoảng loạn cho xã hội như việc loan báo có người đầu tiên tử vong vì virus corona, thúc bách người dân tích trữ lương thực thực phẩm và thuốc men, xuyên tạc về tình hình dịch bệnh COVID-19 tại Việt Nam, công tác phòng chống dịch của Đảng, Nhà nước.

Theo một báo cáo của cơ quan công an, lợi dụng tình hình dịch bệnh với nhiều diễn biến gây lo ngại, nhiều đối tượng thù địch đã thiết lập hàng trăm website và hàng nghìn nhóm (kín và công khai), tài khoản mạng xã hội (Facebook, YouTube, Zalo…) tán phát thông tin dưới nhiều thủ đoạn khác nhau.

Tin giả không chỉ tồn tại trên mạng xã hội mà trên cả các trang báo. Một loạt trang web tin tức như Toronto Sun, New York Post, U.S. Sun và Vietnam Insider, mới đây đăng bài khẳng định mèo đen ở Việt Nam đang bị “thu gom, giết và nghiền thành thuốc chữa virus corona.” Các website này đều trích nguồn từ hãng tin South West News Service, tự nhận là một hãng tin tức và nội dung truyền thông có trụ sở ở Anh và Mỹ với lịch sử hơn 40 năm.

Tác hại của tin giả nhất là tin xuyên tạc được tạo dựng, tán phát từ các cá nhân hay tổ chức trên mạng xã hội liên quan đến các lĩnh vực chính trị, kinh tế, văn hóa, xã hội, quốc phòng, an ninh, đối ngoại,... đều gây nên những hệ quả nghiêm trọng, ảnh hưởng xấu đến đạo đức, lối sống, nhân cách của cá nhân, cộng đồng và văn hóa dân tộc cũng như thái độ, suy nghĩ và quyết định của người dân với nhiều vấn đề của đất nước. Một trong những hệ quả không hề nhỏ mà các tin tức giả gây ra, đó là làm suy giảm niềm tin của công chúng vào truyền thông của đất nước nói chung và báo chí nói riêng, khiến cho công chúng không xác định được đâu là những nguồn tin đáng tin cậy để tiếp nhận... Đáng lưu ý, trong điều kiện các thế lực thù địch, cơ hội, phản động đang triệt để lợi dụng internet và mạng xã hội để tiến hành “diễn biến hòa bình” chống phá cách mạng Việt Nam, điều này không chỉ làm gia tăng số lượng tin giả mà tính nguy hại của nó cũng tăng lên...

## Giải pháp

Khi có quá nhiều tin giả tràn lan trên mạng, vậy sẽ cần một nơi để đưa nguồn tin đúng đắn về phía người dùng. Việc đưa nguồn tin đúng đắn tới các độc giả là điều quan trọng hơn bao giờ hết. Mặc dù có rất nhiều các trang mạng uy tín đã xuất hiênj từ rất lâu nhưng do một vài yếu tố hay do các nền tảng xã hội phát triển lớn mạnh. Nên các trang web đó không thể tới được với đọc giả. Nên tôi đã thực hiện đề tài nhằm mục đích thu thập các thông tin đúng đắn từ nguồn tin tin cậy và thuật toán để đánh giá một bài báo có độ tin cây là bao nhiêu dựa vào các nguồn tin thu thập được.

Việc xây dựng website để cho quản trị viên có thể đánh giá tin tức là tin cây. Giúp cho các độc giả không bị nhiễu vì tin giả tràn lan trên mạng.

## Các vấn đề an toàn

Đối với các vấn đề an toàn của website sẽ phụ thuộc vào độ ổn định của việc crawling dữ liệu. Khi crawling dữ liệu cần đảm bảo tính độc lập không ảnh hưởng tới hệ thống chính. Việc quản lý phiên làm việc của admin để tránh sự tấn công từ hacker nhằm chiếm quyền điều khiển. Do website cung cấp chức năng tìm kiếm và đăng nhập do đó cũng cần phải phòng tránh các lỗi liên quan tới sql injection, reflected xss

Đối với việc crawling dữ liệu vì là thu thập dữ liệu từ website khác cần đảm bảo về các ảnh hưởng từ website đó ví dụ như stored xss. Khi crawling cũng cần sử dụng proxy để tránh bị block IP không thể truy cập vào website

## Kết luận

Đồ án được thực hiện nhằm mục đích tìm hiểu việc thu thập tổng hợp các tin tức từ các nguồn chính xác và cá vấn đề an toàn khi thu thập thông tin việc phát hiển và kiểm tra bài báo là tin thật hay giả dựa trên các nguồn tin đáng tin cậy.

Mục tiêu đặt ra khi thực hiện đồ án là:

1. Thực hiện thu thập thông tin từ các nguồn tin tin cậy
2. Triển khai trang web quản trị và hiển thị bài thông tin thu thập
3. Triển khai các biện pháp an toàn cho website

Sau thời gian thực hiện đồ án, các mục tiêu về cơ bản đã đạt được. Tuy nhiên việc phân tích tin giả là vấn đề khá phức tạp, thời gian thực hiện đồ án tương đối ngắn nên chắc chắn không tránh khỏi thiếu sót. Rất mong được sự góp ý của các thầy cô, cũng như các bạn học viên để đồ án này được hoàn thiện hơn.

|  |  |
| --- | --- |
|  |  |

# CHƯƠNG 2: KHẢO SÁT VÀ PHÂN TÍCH HỆ THỐNG

## 1.PHÂN TÍCH YÊU CẦU

### 1.1. Chức năng

Đây là hệ thống thu thập dữ liệu và phân tích dữ liệu nhằm mục đích chia sẻ dữ liệu tin cậy và kết hợp với các hệ thống khác tạo thành hệ sinh thái.

* Cho phép user các trang web khác cùng truy thực hiện login và sử dụng tài nguyên chung
* Thu thập và phân tích dữ liệu từ các trang web tin cậy. Từ đó có thể phân tích và chuẩn hóa dữ liệu nhằm phục vụ cho mục đích tìm kiếm nhanh và tiếp cận dễ dàng cho người dung
* Cho phép người dùng bình luận các bài báo hiển thị.
* Hiện thị bài báo dựa trên hành vi người dùng.
* Quản trị viên có thể cấu hình cho phép hệ thống thu thập thông các trang web khác

### 1.2. Yêu cầu đặt ra

* Hệ thống sẽ gồm 3 phần

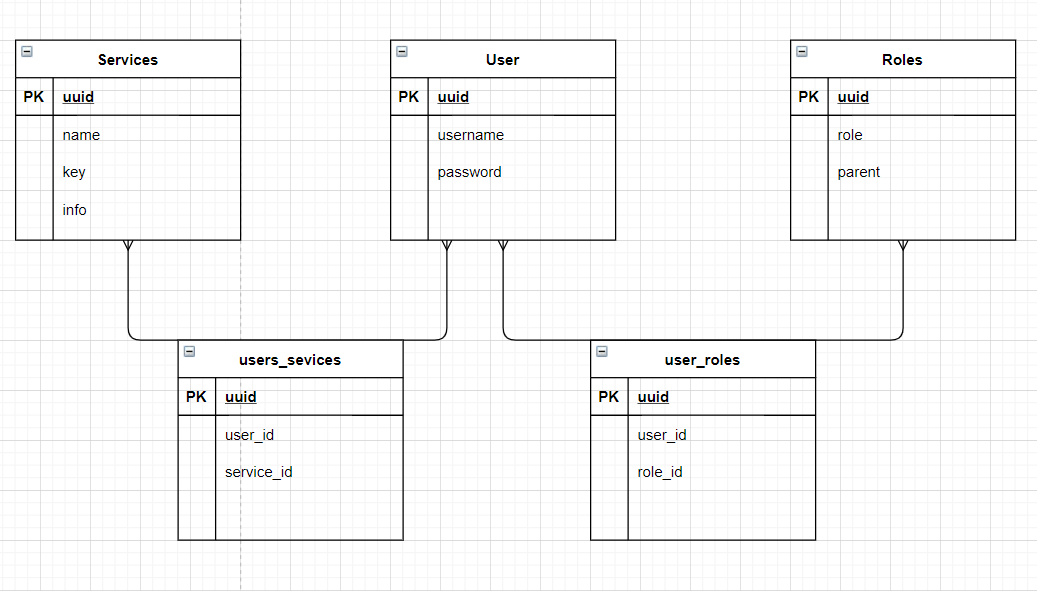
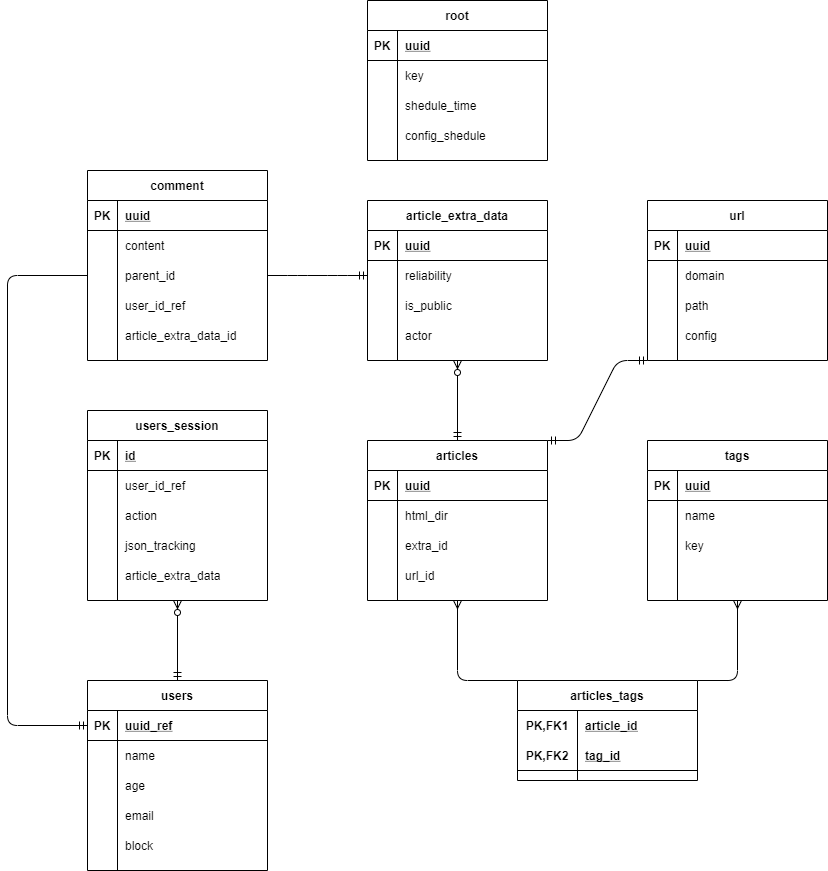
1. Phần hiển thị tin tức trực tuyến:
   1. Hiện thị các bài báo dựa vào sở thích đọc tin tức của người dùng thông qua việc sử dụng trang web
   2. Chức năng tìm kiếm bài bào dựa vào các tiêu chí
   3. Chức năng bình luận và đánh giá chất lượng bài báo
2. Phần Thu thập thông tin
   1. Quản trị viên sẽ thiết kế được việc thu thập thông tin tại các trang web nào
   2. Quan trị viên sẽ theo dõi đươc qua trình thu thập thông tin
   3. Hệ thống sẽ trải qua các bước để thu thập thông tin: tải mã nguồn, phân tích bóc tách mã nguồn, lưu trữ mã nguồn
3. Phân xác thực
   1. Thực hiện xác thực để kết hợp với các dịch vụ khác. Nhằm tạo nên 1 hệ sinh thái an toàn

## 2. PHÂN TÍCH HỆ THỐNG

### 2.1 Biểu đồ tổng quát hệ thống

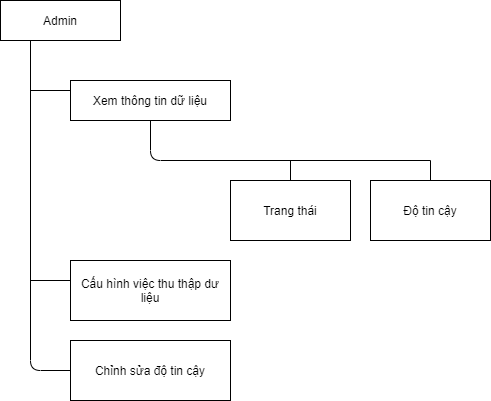
Hình 2.2.1: Sơ đồ thiết kế hệ thống

### 2.2 Cơ sở dữ liệu

1. Cơ sở dữ liệu của AuthenService  
   Hình 2.2.1: Cơ sở dữ liệu của Authen Service
2. Cơ sở dữ liệu của User và thông tin thu thập  
   Hình 2.2.2: Cơ sở dữ liệu của User và thông tin thu thập

### 2.3 Hệ thống thu thập thông tin

1. Sơ đồ phân cấp chức năng



Hình 2.2.3: Sơ đồ phân cấp chức năng hệ thống thu thập thông tin

Chức năng: Xem thông tin dữ liệu

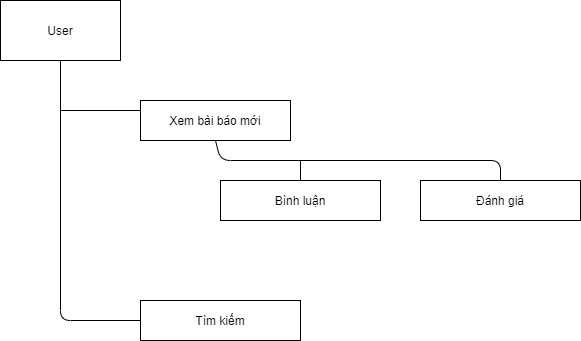
* Dữ liệu sẽ tự động hóa hằng ngày để cập nhập tin tức mới nhất dựa vào các cấu hình có sẵn. Hệ thống sẽ tự đánh giá độ tin cây theo 2 tiêu chí: dựa vào các trang web uy tín do admin lựa chọn ( dantri.com, 24h.com) và dựa vào nội dung của bài báo
* Đầu vào: Ngày tháng
* Các quá trình xử lý
  + Hệ thống sẽ thu thập các thông tin về bài viết mới ở trong cơ sở dữ liệu và hiển thị
* Đầu ra: các bài viết dựa vào ngày tháng.

Chức năng: Cấu hình thu thập dữ liệu

* Bộ selector sẽ đc quản trị viên setup nhằm phục vụ cho việc thu thập dữ liệu.
* Đầu vào: Selector, url
* Các quá trình xử lý
  + Hệ thống sẽ truyền lệnh tới module crawling và đưa thông tin về cấu hình, địa chỉ truy cập
  + Module dựa vào các thông tin nhận được tiến hành thu thập và phân tích dữ liệu sau đó lưu vào trong cơ sở dữ liệu theo ngày, hoặc theo giờ.
  + Các Module crawling chia thành nhiều module nhỏ và hoạt động độc lập ko liên quan tới nhau. Sẽ không ảnh hưởng gì tới người dùng.
* Đầu ra: Thông tin đã qua xử lý trong cơ sở dữ liệu ( nội dung, tiêu đề, loại, độ tin cậy)

### 2.4 Hệ thống hiển thị dữ liệu cho người dùng

1. Sơ đồ phân cấp chức năng



Hình 2.2.4: Sơ đồ phân cấp chức năng hiển thị tin tức cho người dùng

Chức năng: Xem thông tin dữ liệu

* Hệ thống sẽ hiển thị menu cho người dùng lựa chọn và đồng thời hiện chi tiết từng bài báo
* Đầu vào:
* Các quá trình xử lý
  + Hệ thống sẽ hiển thị thông tin bài bào dựa vào thông tin người dùng
* Đầu ra: các bài viết

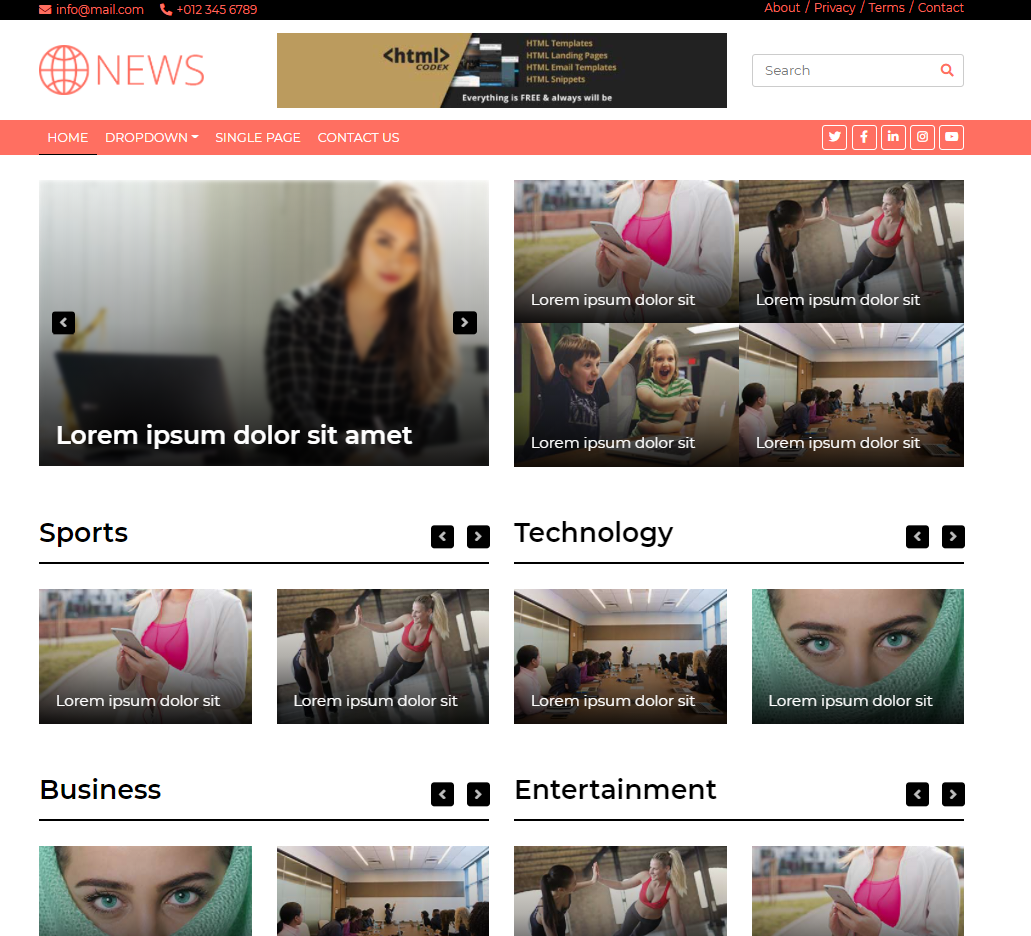
Chức năng: Tìm kiếm

* Dựa vào từ khóa người dùng cung cấp sẽ tìm kiếm trong nội dung bài báo hoặc tiêu đề. Sử dụng cơ sở dữ liệu elasticsearch 1 cơ sở dữ liệu dùng để truy vấn dữ liệu realtime nhanh chóng
* Đầu vào: từ khóa
* Các quá trình xử lý
  + Bóc tách từ khóa và thực hiện multiple search trong tiêu đề và nội dung
* Đầu ra: Các bài báo mà có thể liên quan tới từ khóa

# CHƯƠNG 3: THỰC HIỆN VÀ TRIỂN KHAI HỆ THỐNG WEB TIN TỨC

## 1. Giao diện triển khai

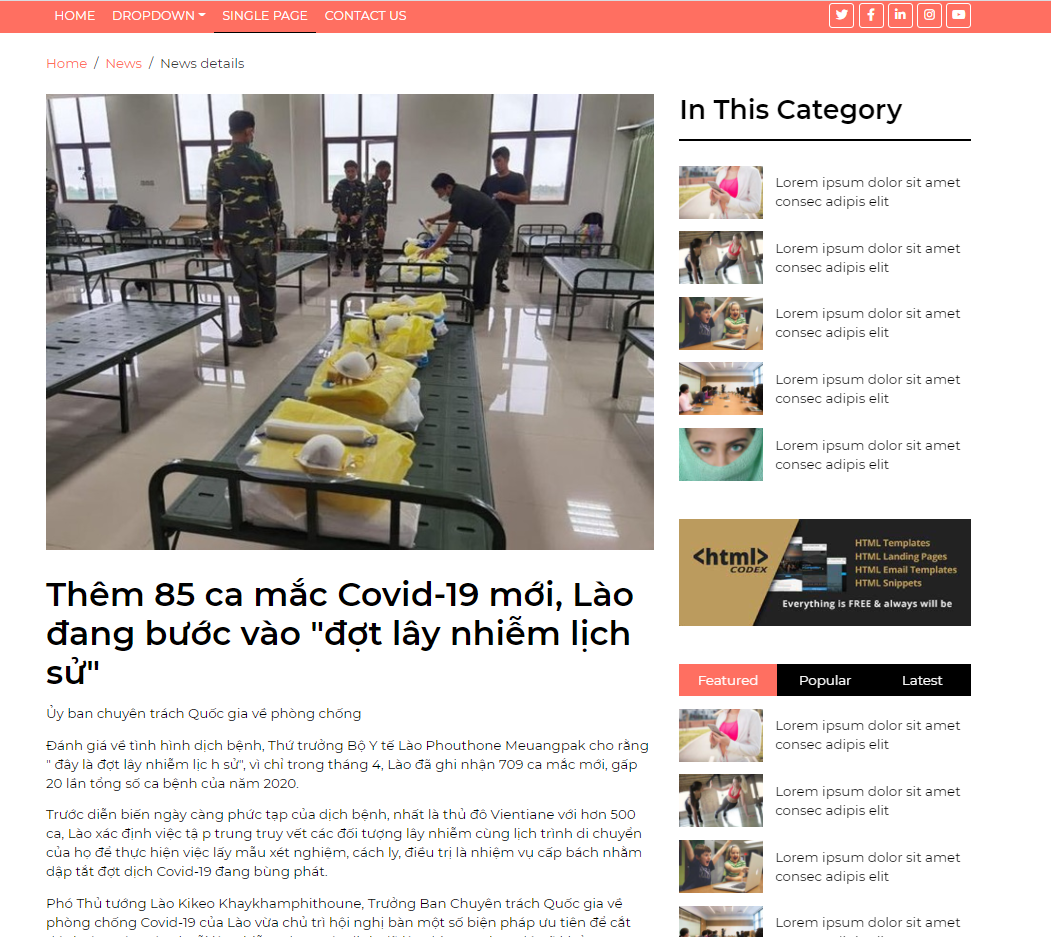
### 1.1 Giao diện trang chủ



Hình 3.1: Giao diện trang chủ người dùng

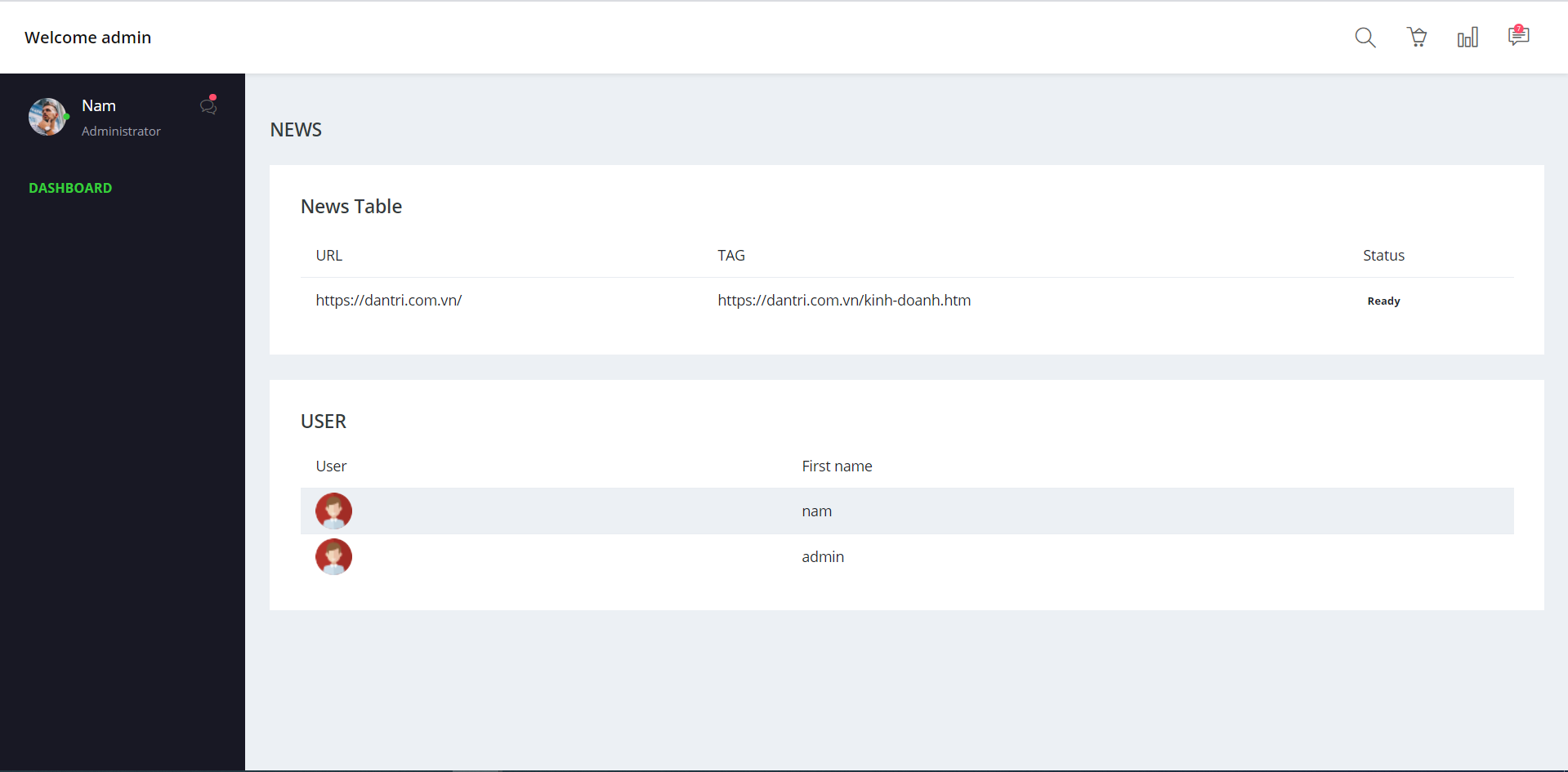
Người dùng thực hiện được tìm kiểm và đọc các bài báo mới

### 1.2 Giao diện bài báo chi tiết



Hình 3.2: Giao diện bài báo chi tiết người dùng

### 1.3 Giao diện admin



Hình 3.3: Giao diện quản trị viên

# CHƯƠNG 4. TRIỂN KHAI WEB CRAWLING

## 1. Giới thiệu

Thiết kế điển hình của các công cụ tìm kiếm là một "tầng", trong đó quá trình thu thập thông tin Web tạo ra một bộ sưu tập được chỉ mục và tìm kiếm. Hầu hết các thiết kế của công cụ tìm kiếm coi việc thu thập thông tin web chỉ là giai đoạn đầu tiên trong tìm kiếm trên web. Đây là một mô hình phân tầng, trong đó các hoạt động được thực hiện theo thứ tự nghiêm ngặt: đầu tiên thu thập thông tin, sau đó lập chỉ mục và sau đó tìm kiếm

Tạo ra web crawling nhằm mục đích tìm kiếm thông tin trên web một cách hiệu quả. Việc tìm kiếm thông tin trên Internet là vấn đề quan trọng.

## 2. Đặc điểm website

Một trong những việc khó khăn là định hình cấu trúc website. Vì mỗi website đều có cấu trúc khác nhau và lượng thông tin khác nhau

Có hai phương thức chính để lấy về các mẫu của trang web:

Lấy mẫu dọc: liên quan đến việc thu thập các trang bị hạn chế bởi tên miền. Vì Khi hệ thống tên miền tạo ra cấu trúc phân cấp, việc lấy mẫu dọc có thể được thực hiện ở các cấp độ khác nhau của cấu trúc. Khi lấy mẫu dọc được thực hiện ở cấp cao nhất, nó có thể chọn toàn bộ các quốc gia, chẳng hạn như .cl, .it, .au, dự kiến sẽ gắn kết về ngôn ngữ, chủ đề, lịch sử hoặc có thể chọn các miền cấp cao nhất chung như .edu hoặc .com, ngoại trừ tên miền .gov. Khi lấy mẫu dọc được thực hiện ở cấp độ thứ hai, nó sẽ chọn một tập hợp các trang do các thành viên của cùng một tổ chức tạo ra ví dụ actvn.edu.vn.

Lấy mẫu theo chiều ngang: liên quan đến tiêu chí lựa chọn không dựa trên tên miền. Trong trường hợp này sẽ có hai cách để thu thập dữ liệu: sử dụng nhật ký của các phiên truyền dữ liệu trong proxy của một tổ chức lớn hoặc thông qua ISP, hăọc sử dụng web crawler. Có những ưu điểm và nhược điểm cho từng phương pháp: khi chúng ta lắng nghe một proxy sẽ dễ dàng tìm kiếm được các trang web phổ biến. nhưng số lần truy cập lại khá là khó kiểm soát, điều đó lại phụ thuộc vào người dùng. Sử dụng web crawler mức độ phổ biến của các trang pphải được ước tính nhưng khoảng thời gian truy cập lại có thể tùy chỉnh.

Trong việc lấy mẫu theo chiều ngang, việc truy cập ngẫu nhiên có thể sử dụng để có được một tập hợp các trang gần như được truy cập với xác suất tỷ lệ thuận với giá trị thứ hạng của trang web và sau đó việc lấy mẫu sẽ ddược lấy từ tập hợp này với xác suất tỷ lệ nghịch với thứ hạng của trang web vì vậy những mẫu được mong đợi sẽ không bị sai lệch.

## 3. Web sẽ có sự thay đổi thường xuyên

Thì website sẽ có sự thay đổi về trạng thái như thêm mới page, xóa page, hay cập nhật pages. Việc thực hiện cập nhật dữ liệu được thực hiện bằng cách truy cập vào nhiều trang web trong một khoảng thời gian ngắn. Đối với mỗi lần truy cập thông tin sau đây sẽ luôn luôn có sẵn:

* Thời gian truy cập vào trang web
* Lần cuối thay đổi của trang web
* Sử dụng văn bản cũ đã được thay thu thập để tìm kiếm sự thay đổi của trang web

Các thông tin dưới đây có thể ước tính nếu thời gian thu thập cách nhau ngắn:

* Thời điểm trang web xuất hiện lần đầu tiên
* Thời điểm mà trang web không thể truy cập được nữa

Trong mọi trường hợp, kết quả chỉ là ước tính của các giá trị thực tế vì chúng có được bằng cách thu thập thông tin và phát hiện dựa trên sự thay đổi chứ không phải bởi các sự kiện thông báo khác vì vậy giữa hai lần truy cập một trang web sẽ có sự tha đổi rất nhiều.

Các đặc điểm thay đổi trên trang web:

* Thời gian tồn tại website: thời gian xóa – thời gian lần đầu xuất hiện
* Số lần thay đổi của trang web trong suốt gian tồn tại
* Khoảng thời giant hay đổi trung bình của website: thời gian tồn tại chia cho số lần thay đổi

## 4. Lập chỉ mục và truy vấn các web

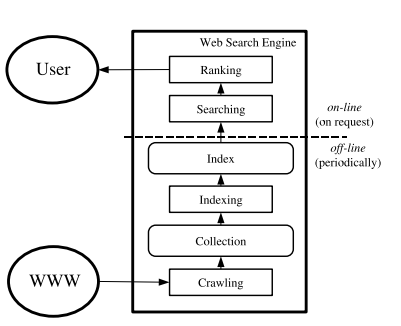
Quá trình thu thập tìm kiếm trên web sẽ có hai phần chính: ngoại tuyến và trực tuyến.

* Việc thu thập ngoại tuyến được thực thi định kì bời công cụ tìm kiếm và bao gồm việc tải xuống một tập hợp của web để xây dựng một tập hợp các trang, sau đó chuyển đổi thành một chỉ mục có thể tìm kiếm được
* Việc thu thâpj trực tuyến được thực thi mỗi khi thực thi truy vấn người dùng và sử dụng chỉ mục để chọn một số tài liệu ứng viên được sắp xếp theo ước tính về mức độ liên quan của chúng đối với nhu cầu người dùng.

Quá trình được thể hiện trong hình 3.6

Mỗi một trang web sẽ có rất nhiều định dạng khác nhau như văn bản thuần túy, hay mã code HTML, hay tài liệu PDF, CSV hay hình ảnh PNG, JPEG và các định dạng khác. Việc đầu tiên để thiết lập chỉ mục của trang web là trích xuất một chế độ xem logic từ các tài liệu chuẩn. Thường thấy chế độ tìm kiếm logic nhiều nhất của người trên trang web là các từ khóa. Đối với mỗi tài liệu chỉ được xem như một tập hợp các từ không có thứ tự. Trong các công cụ tìm kiếm web hiện đại, chế độ xem này được mở rộng với thông tin bổ sung liên quan đến tần số từ và các thuộc tính định dạng văn bản, cũng như các thông tin mở rộng về các trang web bao gồm các mô tả được nhúng và các từ khóa rõ rang trong HTML đánh dấu.

Có một số hoạt động để chuẩn hóa các văn bản được thực hiện để trích xuất các từ khóa, những cái được sử dụng nhiều nhất là: mã hóa, loại bỏ từ khóa và tạo gốc.



Hình 3.6: quá trình web thực hiện tạo chỉ mục

### 5. Các lỗi về web crawling

Có hai đặc điểm quan trọng của Web tạo ra một kịch bản trong đó thu thập dữ liệu Web rất khó khăn: khối lượng lớn và tốc độ thay đổi của nó, vì có một lượng lớn các trang được thêm vào, đã thay đổi và loại bỏ mỗi ngày. Ngoài ra, tốc độ mạng đã được cải thiện ít hơn so với tốc độ xử lý và lưu trữ hiện tại năng lực. Khối lượng lớn ngụ ý rằng trình thu thập thông tin chỉ có thể tải xuống một phần nhỏ các trang Web trong một thời gian nhất định, vì vậy nó cần ưu tiên các lượt tải xuống. Tỷ lệ thay đổi cao ngụ ý rằng vào thời điểm trình thu thập thông tin đang tải xuống các trang cuối cùng từ một trang web, rất có thể các trang mới đã được thêm vào trang web, hoặc các trang đã được cập nhật hoặc thậm chí đã bị xóa.

Thu thập thông tin trên Web, theo một cách nào đó, giống như xem bầu trời trong một đêm quang đãng: những gì chúng ta thấy phản ánh trạng thái của các ngôi sao tại các thời điểm khác nhau, khi ánh sáng truyền đi những khoảng cách khác nhau. Những gì trình thu thập thông tin web nhận được không phải là "snapshot" của Web, vì nó không đại diện cho Web tại bất kỳ thời điểm nhất định nào. Các trang cuối cùng được thu thập thông tin có thể được trình bày rất chính xác, nhưng các trang đầu tiên được tải xuống có khả năng cao là đã được thay đổi

Hoạt động của trình thu thập thông tin Web là kết quả của sự kết hợp các chính sách:

* Chính sách lựa chọn cho biết trang nào sẽ tải xuống.
* Chính sách truy cập lại nêu rõ khi nào cần kiểm tra các thay đổi đối với các trang.
* Chính sách lịch sự quy định cách tránh làm quá tải các trang Web.
* Chính sách song song nêu rõ cách điều phối các trình thu thập thông tin Web phân tán.

### 5.1 Chính sách tuyển chọn

Với kích thước hiện tại của Web, ngay cả các công cụ tìm kiếm lớn cũng chỉ bao gồm một phần của nội dung. Vì một trình thu thập thông tin luôn tải xuống chỉ một phần nhỏ các trang Web, phần tải xuống rất mong muốn phần tải xuống chứa các trang liên quan nhất, chứ không phải một trang web cụ thể.

Điều này đòi hỏi một số liệu về mức độ quan trọng để ưu tiên các trang Web. Tầm quan trọng của một trang là một chức năng về chất lượng nội tại của nó, mức độ phổ biến của nó về các liên kết hoặc lượt truy cập, và thậm chí cả URL của nó (sau này là trường hợp các công cụ tìm kiếm dọc bị hạn chế trong một tên miền cấp cao nhất hoặc các công cụ tìm kiếm bị hạn chế đến một cố định trang mạng). Việc thiết kế một chính sách lựa chọn tốt có thêm một khó khăn: nó phải làm việc với một phần thông tin, vì tập hợp đầy đủ các trang web không được xác định trong quá trình thu thập thông tin.

Tầm quan trọng của một trang đối với trình thu thập thông tin cũng có thể được thể hiện dưới dạng một hàm về mức độ giống nhau của một trang. Vấn đề chính trong thu thập thông tin tập trung là trong ngữ cảnh của trình thu thập thông tin Web, ta có thể dự đoán sự giống nhau của văn bản của một trang nhất định với truy vấn trước khi thực sự tải xuống. Hiệu suất của việc thu thập thông tin tập trung chủ yếu phụ thuộc vào mức độ phong phú của các liên kết trong chủ đề cụ thể đang tìm kiếm và thu thập thông tin tập trung thường dựa vào một công cụ tìm kiếm web chung để cung cấp các điểm xuất phát.

### 5.2 Chính sách về tái cập nhật trang web

Trang web luôn luôn biến động và việc thu thập một phần thông tin của mọt trang web có thể mất nhiều thời gian thường được đo bằng tuần hoặc tháng. Vào thời điểm thu thập thông tin của trang web hoàn thành việc thu thập thông tin, nhiều sự kiện có thể đã xảy ra. Chúng mô tả các sự kiện là tạo mới, cập nhật và xóa bỏ

### 5.3 Chính sách chạy song song

Trình thu thập thông tin song song là trình thu thập thông tin chạy song song nhiều quá trình. Mục tiêu là tối đa hóa lượt tải xuống đồng thời giảm thiểu chi phí từ quá trình song song hóa và tránh tải xuống lặp lại cùng một trang. Để tránh tải xuống cùng một trang nhiều lần, hệ thống thu thập thông tin yêu cầu một chính sách để chỉ định các URL mới được phát hiện trong quá trình thu thập thông tin, vì cùng một URL có thể được tìm thấy bởi hai quy trình thu thập thông tin.

## 6. Kiến trúc web crawler

Một Crawler cũng phải có quy trình chiến lược thu thập thông tin tốt nhất. nó cần kiến trúc để tối ưu hóa việc thu thập. Việc thu thập thông tin cũng là một phần của công cụ tìm kiếm. Khi bản thiết kế Crawler được công bố thường thiếu chi tiết quan trọng điều đó ngăn cản người khác sao chép tác phẩm.

Một Crawler sẽ phải đối phó với nhiều thách thức cùng một lúc, và đôi lúc các crawler sẽ xảy ra xung đột như crawler A đã chạy nhưng crawler vẫn tiếp tục chạy thâm chí còn ghi đè kết quả đôi lúc sẽ gây sai lệch về tính toán. Chúng phải giữ lại các trang web và phải truy cập lại và trong lúc đó chúng cũng phải cập nhật những trang web mới mà chúng tìm thấy trong các trang đã sửa đổi. Hầu hết khi sử dụng crawler sẽ chiếm dụng hết các tài nguyên như bang thông mạng và cũng không được phép làm quá tải web mà chúng truy cập tới.

### 6.1. Các vấn đề liên quan tới lập lịch

Tại đây một Crawler cần tải xuống một tập hợp các trang web, mới mỗi trang p có kích thước S­­p được đo bằng byte sử dụng kết nối mạng có dung lượng B tính bằng byte trên giây. Mục tiêu cần thu thập thông tin là tải xuống tất cả các trang trong thời gian nhỏ nhất. Giải pháp nhỏ cho vấn đề này là tải xuống tất cả các trang web đồng thời và đối với mỗi trang sử dụng bang thông tỷ lệ với kích thước của mỗi trang.

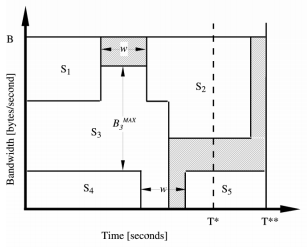
Nếu Bp là tốc độ tải xuống của trang một trang web p vậy thì:

Với T là thời gian tối ưu để sử dụng tất cả các băng thông có sẵn

Tuy nhiên có nhiều hạn chế ngăn cản có thể chạy đúng quá trình một cách suôn sẻ. Một trong những số đó là chính sách lập lịch phải tránh làm quả tải các trang web, Crawler không nên tải xuống nhiều hơn một trang từ một trang web cùng một lúc nó phải đợi vài giây giữa các lần yêu cầu.

Thay vì tải xuống tất cả các trang song song, chúng tôi cũng có thể tuần tự hóa tất cả các yêu cầu, chỉ tải xuống mỗi lần một trang ở tốc độ tối đa. Tuy nhiên băng thông có sẵn cho các trang web BiMAX thường thấp hơn băng thông trình thu thập thông tin vì vậy kịch bản này không thực tế.

Trong thực tế các đường thời gian tải xuống sẽ tương tự như đường biểu diễn trong Hình 3.7. Thời gian tối ưu T\* không đạt được, vì một số băng thông bị lãng phí do hạn chế về tốc độ của các trang web, và để thực tế là Crawler sẽ phải đợi giữa các lần truy cập vào một trang web.



Hình 3.7: Dòng thời gian tải xuống thực tế trong quá trình thu thập Các trang 1−2 và 4−5 thuộc về đến cùng một trang web và trình thu thập thông tin sẽ đợi w giây giữa chúng. Phần đã nở bị lãng phí băng thông do các ràng buộc trong chính sách lập lịch trình. Thời gian tối ưu T\* không đạt được.

Để khắc phục các vấn đề rõ ràng là chugns ta nên cố gắng bão hòa mạng liên kết, tải các trang từ nhiều trang web khác nhau cùng một lúc. Thật không may, hầu hết các trang nằm trong một số lượng nhỏ các trang web: sự phân phối các trang đến các trang web điều đó sẽ thật tệ trong quá trình mở rộng Crawler. Do đó, không thể sử dụng một cách hiệu quả một số lưuọng lớn các file robos của nó khó đạt đưojc viêcj sử dụng cáo băng thông có sẵn.

Có một hạn chế nghiêm trọng khác: yêu cầu HTTP sẽ có độ trễ, và thời gian chờ có thể là trên 25% tổng thời gian của yêu cầu. Độ trễ này chủ yếu là thời gian cần thiết để thiết lập kết nối TCP và nó có thể khắc phục một phần nếu cùng một kết nối được sửa dụng để đưa ra một số yêu cầu bằng cách sử dụng tính năng “keep-alive” của HTTP/1.1

### 6.2. Các vấn đề liên quan tới mô hình hoạt động

Việc thu thập được biểu đạt thông qua 2 từ “crawler” và “spider”, và những từ đó đều tư tưởng về một đồ thị có hướng. Điều đó khác rất xa so với những gì đang thực sự xảy ra vì thu thập thông tin chỉ tự động tải xuống trang không cần tuân theo kiểu duyệt web trong một số trường hợp phương pháp tiếp cân theo chiều rộng sẽ được sử dụng, trong các trường hợp khác, việc thu thập thông tin được thực hiện theo cách không thể hiện rõ ràng trên biểu đồ Web, và không giống với đường truyền biểu đồ.

Thuật toán thu thập thông tin điển hình xuất hiện từ những ngày đầu của World Wide Web, và họ được đưa ra bởi thuật toán:

* Xác định Q = {p1, p2, …, pn} với p là các page của URL
* Khởi tạo V = 0, đi tới các URLs
* Cho tới khi Q rỗng thực hiện
  + Gán V = V + {p}
  + Tải p về là tìm kiếm các đường dẫn p’
  + tìm kiếm p’ sao cho p’ không thuộc V và không thuộc q thêm vào Q

Đối với thuật toán này ta có thể cải thiện, vì trong quá trình thu thập thông tin, không cần thiết phải thêm mới các URL được tìm thấy tới Q mỗi khi trang web được phân tích và bóc tách. Nhưng URL mới sẽ được thêm vào một nhóm tập hợp khác.

**Việc lập chỉ mục được thực hiện theo số lượng lớn**. Quá trình thu thập thông tin thêm thông tin vào một bộ sưu tập sẽ được lập chỉ mục. Quá trình lập chỉ mục được thực hiện hàng loạt, nhiều megabyte văn bản cùng một lúc và với các thuật toán hiện tại sẽ rất kém hiệu quả nếu làm từng tài liệu một, trừ khi người ta có thể đạt được sự cân bằng chính xác giữa luồng tài liệu đến và tốc độ xử lý của chỉ mục, và trong trường hợp này, việc xây dựng chỉ mục sẽ trở thành điểm nghẽn. Do đó, trong hầu hết các công cụ tìm kiếm, chỉ mục không được cập nhật liên tục nhưng hoàn toàn cùng một lúc.

**Trình thu thập thông tin được phân phối trao đổi URL theo số lượng lớn.** Nếu trình thu thập thông tin được phân phối, thì nó phải gửi trình thu thập thông tin trở lại máy chủ trung tâm hoặc phải trao đổi kết quả với các quy trình thu thập thông tin khác. Để hiệu suất tốt hơn, nó sẽ phải gửi nhiều URL cùng một lúc, vì việc trao đổi URL sẽ tạo ra chi phí hầu hết được cũng cấp khi chuyển đổi nội dung. Điều này có nghĩa là cách các URL được sắp xếp cục bộ sẽ không ảnh hưởng đến thứ tự thu thập thông tin.

**Các URL quan trọng được xem trước đó trong thu thập thông tin.** Nếu một số thứ tự URL được thực hiện và nếu thứ tự này là không dựa trên sự giống văn bản với một truy vấn, thì ở trạng thái ổn định, một trang mà chúng ta vừa thấy là ứng cử viên khó có thể được tải xuống trong tương lai gần: Các trang "tốt" được nhìn thấy sớm trong quá trình thu thập thông tin quá trình. Ngược lại, nếu URL được nhìn thấy lần đầu tiên trong giai đoạn cuối của quá trình thu thập thông tin, có khả năng cao đó không phải là một trang rất thú vị. Điều này rõ ràng là đúng nếu độ ưu tiên của trang web được áp dụng, bởi vì nó phản ánh thời gian một người lướt ngẫu nhiên "dành" trên trang và nếu ngẫu nhiên người lướt sóng dành nhiều thời gian hơn trong một trang, thì có thể trang đó có thể được truy cập từ một số liên kết.

Tôi nhận thấy rằng công việc trước đây có xu hướng tách biệt hai vấn đề giống nhau và trộn lẫn hai vấn đề khác nhau các vấn đề:

* Hai vấn đề khác nhau thường được trộn lẫn là vấn đề hiệu quả ngắn hạn (tối đa hóa việc sử dụng băng thông và lịch sự với máy chủ) và hiệu quả dài hạn (ra lệnh thu thập thông tin quy trình để tải các trang quan trọng trước). Chúng tôi thảo luận tại sao hai vấn đề này có thể được tách biệt trong
* Hai vấn đề liên quan thường được coi là các vấn đề riêng biệt là làm mới chỉ mục và chỉ số chất lượng nội tại. Chúng tôi cho rằng tốt hơn là nên suy nghĩ về một loạt các điểm số liên quan đối với các đặc điểm khác nhau của các tài liệu trong bộ sưu tập, bao gồm cả độ mới, có trọng số tương ứng với một số mức độ ưu tiên khác nhau tùy thuộc vào ngữ cảnh sử dụng của trình thu thập thông tin.

### 6.3. Tách biệt ngắn hạn với lập kế hoạch dài hạn

Để giải quyết riêng việc lập lịch dài hạn và lập lịch ngắn hạn. Để có thể làm được điều này, ta phải chứng minh rằng cả hai vấn đề có thể được tách biệt, cụ thể là ta phải kiểm tra xem chất lượng nội tại của một trang Web hoặc máy chủ Web có liên quan đến băng thông có sẵn để tải xuống trang đó. Nếu đúng như vậy, thì tôi sẽ không thể chọn những trang quan trọng nhất trước tiên và sau đó sắp xếp lại các trang để sử dụng băng thông hiệu quả, bởi vì trong khi chọn các trang quan trọng, chúng tôi sẽ ảnh hưởng đến tốc độ truyền mạng.

### 6.4. Kết hợp chất lượng và độ mới của trang

Trình thu thập thông tin của công cụ tìm kiếm được thiết kế để tạo một tập hợp các trang hữu ích cho chỉ mục của công cụ tìm kiếm. Để hữu ích, chỉ số này nên cân bằng giữa tính toàn diện và chất lượng. Hai mục tiêu này cạnh tranh, bởi vì ở mỗi bước lập lịch trình, trình thu thập thông tin phải quyết định giữa việc tải xuống một trang mới, không phải hiện tại được lập chỉ mục hoặc làm mới một trang có thể đã lỗi thời trong chỉ mục. Có sự đánh đổi giữa số lượng (nhiều đối tượng hơn) và chất lượng (nhiều đối tượng cập nhật hơn).

Sau đây, tôi đề xuất một chức năng để đo lường chất lượng của chỉ mục của một công cụ tìm kiếm. Các mục tiêu của trình thu thập thông tin là tối đa hóa chức năng này.

Tôi bắt đầu bằng cách nêu ba yếu tố có liên quan đến chất lượng của một trang Web trong chỉ mục:

**Chất lượng nội dung của trang.** Chỉ mục phải chứa một số lượng lớn các trang web thú vị cho người dùng của công cụ tìm kiếm. Tuy nhiên, định nghĩa về những gì sẽ thú vị đối với người dùng là một định nghĩa khá trơn và hiện đang là một chủ đề được nghiên cứu rất nhiều. Một số chiến lược đã được đề xuất, thường dựa vào một chức năng xếp hạng để sắp xếp danh sách các đối tượng được tìm thấy bởi máy tìm kiếm. Chúng tôi không thể biết trước sự quan tâm mà một trang Web sẽ có đối với người dùng, nhưng chúng ta có thể ước lượng nó sử dụng chức năng xếp hạng xem xét thông tin từng phần mà trình thu thập thông tin đã thu được cho đến nay trong suốt quá trình của nó.

Chất lượng nội tại của một trang có thể được ước tính theo nhiều cách:

* Phân tích liên kết (mức độ phổ biến của liên kết)
* Tương tự với một truy vấn nhất định.
* Quyền truy cập vào trang trên chỉ mục (mức độ phổ biến sử dụng)
* Dựa trên vị trí, theo độ sâu cảm nhận (ví dụ: số lượng thư mục trên đường dẫn đến đối tượng Web).
* Tên miền, địa chỉ IP hoặc phân đoạn, địa lý, v.v.

**Chất lượng đại diện của trang trong chỉ mục.** Mọi đối tượng trong chỉ mục phải đại diện chính xác một đối tượng thực trong Web. Điều này liên quan đến cả lượng dữ liệu được lưu trữ về đối tượng và đến thời gian kết xuất của đối tượng. Chất lượng biểu diễn chủ yếu phụ thuộc vào số lượng và định dạng của thông tin được lưu trữ cho mọi đối tượng. Trong trường hợp của các trang Web, chúng tôi có thể sắp xếp chất lượng đại diện từ ít hơn đến nhiều hơn:

* URL
* URL + Chỉ mục của văn bản gần các liên kết đến trang đó
* URL + Chỉ mục của văn bản gần các liên kết đến trang đó + Chỉ mục của toàn văn
* URL + Chỉ mục của văn bản gần các liên kết đến trang đó + Chỉ mục của toàn văn + Tóm tắt văn bản (“đoạn trích”) được trích xuất bằng cách sử dụng các kỹ thuật xử lý ngôn ngữ tự nhiên hoặc đơn giản bằng cách lấy một vài từ từ đầu văn bản
* URL + Chỉ mục của văn bản trong các liên kết đến trang đó + Chỉ mục của văn bản + tất cả vẵn bản

Có những khả năng khác, chỉ liên quan đến việc lập chỉ mục các phần của trang bằng cách sử dụng đánh dấu HTML như hướng dẫn, tức là chỉ lập chỉ mục tiêu đề, siêu dữ liệu và / hoặc tiêu đề trang.

Thời gian sản xuất phụ thuộc vào định dạng, đặc biệt nếu sử dụng tính năng nén. Một số khả năng thích ứng có thể đã sử dụng ví dụ: văn bản hoặc hình ảnh có thể được nén ngoại trừ những đối tượng yêu cầu chất lượng đại diện lớn vì chúng được công cụ tìm kiếm truy cập thường xuyên

## 7. Thu thập thông tin trên Web Vô hạn

Tôi đã xem xét rằng Web bị giới hạn, nhưng một lượng lớn các trang Web có sẵn công khai được tạo động theo yêu cầu và chứa liên kết đến các trang được tạo động khác. Điều này thường dẫn đến các trang Web có thể được coi là có nhiều trang tùy ý.

Điều này đặt ra một vấn đề đối với việc thu thập thông tin trên Web, vì nó phải được thực hiện theo cách nó ngừng tải xuống các trang từ mỗi trang Web tại một số thời điểm. Nhưng Crawler phải đi sâu bao nhiêu?

Trong mục này sẽ nói về:

* Tôi đề xuất các mô hình để lướt ngẫu nhiên bên trong một trang Web khi số lượng trang không bị giới hạn. Đối với đó, chúng tôi lấy cây do biểu đồ Web của một trang web tạo ra và nghiên cứu nó theo các cấp độ.
* Tôi phân tích các mô hình này, tập trung vào câu hỏi người dùng “sâu” như thế nào bên trong một trang Web.
* Tôi xác nhận các mô hình này bằng cách sử dụng dữ liệu thực tế từ các trang Web, cũng như sử dụng thước đo phân tích liên kết chẳng hạn như thứ hạng của một trang web

### 7.1 Trang web tĩnh và trang web động

Hầu hết các nghiên cứu về Web chỉ đề cập đến "Phần có thể lập chỉ mục công khai", loại trừ một phần của Web đã được gọi "Web ẩn" hay “deep Web”. Không thể lập chỉ mục phần này được mô tả là tất cả các trang mà người dùng bình thường cuối cùng có thể truy cập, nhưng các tác nhân tự động như vì các trình thu thập thông tin được sử dụng bởi các công cụ tìm kiếm không thể.

Một số trang nhất định không thể lập chỉ mục vì chúng yêu cầu đăng ký trước đó hoặc một số ủy quyền đặc biệt chẳng hạn như mật khẩu, hoặc chỉ khả dụng khi được truy cập từ trong một mạng nhất định chẳng hạn như một công ty mạng nội bộ. Những trang khác là các trang động, được tạo sau khi yêu cầu đã được thực hiện. Một số lần họ không

có thể lập chỉ mục vì chúng yêu cầu một số tham số nhất định làm đầu vào, ví dụ: cụm từ truy vấn và các cụm từ truy vấn đó không xác định tại thời điểm thu thập thông tin.

Tuy nhiên, nhiều trang động có thể lập chỉ mục, vì các tham số để tạo chúng có thể được tìm thấy bằng các liên kết sau đây. Đây là trường hợp của, ví dụ: danh mục sản phẩm điển hình trong các cửa hàng trực tuyến, trong đó có các liên kết đến điều hướng danh mục mà người dùng không cần phải đặt ra một truy vấn.

Web thường được coi là một tập hợp các trang, theo nghĩa giống như trong Thông tin truyền thống Bộ sưu tập truy xuất, nhưng lớn hơn nhiều. Theo giả định này, biểu đồ Web có một số lượng hữu hạn các nút trong đó các phép đo như đường kính được xác định rõ ràng. Điều này là sai cơ bản. Lượng thông tin trong Web tại bất kỳ thời điểm nào chắc chắn là hữu hạn, nhưng khi một trang động dẫn đến một trang động khác, số lượng trang có thể là vô hạn. Ví dụ: một trang động triển khai lịch, bạn luôn có thể nhấp vào “tháng tới” và từ một thời điểm nào đó sẽ không còn mục dữ liệu nào nữa trong lịch; con người có thể chắc chắn một cách hợp lý rằng rất khó có thể tìm thấy các sự kiện được lên lịch trước 50 năm, nhưng một trình thu thập thông tin thì không. Trình thu thập thông tin không thể biết khi nào hai trang phản ánh giống nhau thông tin. Có nhiều ví dụ khác về "bẫy trình thu thập thông tin" liên quan đến các vòng lặp và / hoặc các bản sao gần như trùng lặp có thể được phát hiện sau đó, nhưng chúng tôi muốn tránh tải chúng xuống.

Điều này đặt ra một vấn đề đối với việc thu thập thông tin trên Web, vì nó phải được thực hiện theo cách nó ngừng tải xuống các trang từ mỗi trang Web tại một số thời điểm. Hầu hết các nhà nghiên cứu thường áp dụng một trong các cách tiếp cận sau đây để giải quyết vấn đề này:

**Chỉ tải xuống những trang tĩnh.** Phương pháp phỏng đoán thông thường để làm như vậy là tránh tải xuống các URL có chứa dấu ques tion, nhưng phương pháp này có thể không thành công vì có nhiều URL được tạo động nhưng lại không sử dụng tiêu chuẩn CGI, mã hóa các tham số trong phần còn lại của URL. Ngoài ra, một phần có giá trị của các trang Web có sẵn công khai được tạo động theo yêu cầu và không rõ tại sao những các trang sẽ bị phạt vì các trang tĩnh.

**Tải xuống các trang động chỉ với một bộ thông số**. Khi thực hiện việc này, các trang động sẽ được tải xuống với tập hợp các thông số của lần đầu tiên chúng được tìm thấy hoặc với một tập hợp thông số trống. Hạn chế rõ ràng là các trang động có thể truy vấn cơ sở dữ liệu và một tập hợp tham số không thể đại diện cho nội dung của cơ sở dữ liệu.

**Tải xuống số lượng trang tối đa.** Điều này tạo ra một tập dữ liệu phụ thuộc nhiều vào chiến lược nhập thu thập thông tin. Hơn nữa, điều này không thể được sử dụng để so sánh, ví dụ, lượng thông tin trên các miền khác nhau

**Tải xuống một lượng trang nhất định cho mỗi tên miền.** Vì một khoản tiền nhỏ phải trả để đăng ký một tên miền, nên có một số nỗ lực nhất định liên quan đến việc tạo ra một trang Web dưới một tên miền. Tuy nhiên, có một số tên miền nhất định như “.com.vn” rất lớn và có thể yêu cầu các quy tắc đặc biệt.

**Tải xuống đến một lượng cấp độ nhất định** cho mỗi trang Web Bắt đầu từ trang chủ của mỗi trang Web, theo các liên kết đến một độ sâu nhất định. Đây là cách tiếp cận mà chúng tôi xem xét trong bài báo này và tự nhiên câu hỏi là: trình thu thập thông tin phải đi sâu bao nhiêu.

Web của nội dung được tạo động được nhiều trình thu thập dữ liệu Web thu thập thông tin hời hợt, trong một số các trường hợp do trình thu thập thông tin không thể cho biết một URL động từ một URL tĩnh, và trong các trường hợp khác có mục đích.

Tuy nhiên, một số trình thu thập thông tin sẽ đi sâu hơn, trừ khi họ biết khi nào nên dừng và cách xử lý các trang động với các liên kết đến các trang năng động hơn.

Tôi giải quyết vấn đề nắm bắt một phần có liên quan của nội dung được tạo động bằng các thông số đã biết, đồng thời tránh tải quá nhiều trang. Tôi muốn biết liệu người dùng có bao giờ thấy một trang được tạo động hay không. Nếu xác suất quá thấp, công cụ tìm kiếm có muốn lấy lại trang đó? Rõ ràng, từ trang Web hoặc quan điểm của người tìm kiếm, câu trả lời phải là có, nhưng từ quan điểm của công cụ tìm kiếm, câu trả lời có thể là không.

### 7.2 Mô hình người lướt ngẫu nhiên cho một trang web vô hạn

Tôi coi một trang web S = {Page, Links} dưới dạng một tập hợp các trang dưới cùng một tên máy chủ, tạo thành một đồ thị có hướng. Các nút trang Pages= {P1,P2,…} và các vòng cung là các Liên kết sao cho (Pi,Pj) Links nếu có tồn tại liên kết từ trang Pi tới Pj trong trang web.

**Định nghĩa (Phiên người dùng)** Tôi xác định phiên người dùng **u** là một chuỗi hữu hạn các lần xem trang **u** = (P1, P2,…, Pn) với Pi Pages, và (Pi, Pi+1) Links. Yêu cầu đầu tiên u0 không cần phải là trang bắt đầu nằm ở thư mục gốc của máy chủ, vì một số người dùng có thể vào trang Web theo liên kết đến trang nội bộ, ví dụ: nếu chúng đến từ một công cụ tìm kiếm.

**Định nghĩa (Độ sâu trang)** Đối với một trang Pi và một phiên u, chúng tôi xác định độ sâu của trang trong phiên, độ sâu (Pi, **u**) như:

Độ sâu về cơ bản là độ dài của đường dẫn ngắn nhất từ trang bắt đầu đến các trang thực sự nhìn thấy trong một phiên. Lưu ý rằng độ sâu của một trang không chỉ là một chức năng của cấu trúc trang Web, nó còn là độ sâu cảm nhận được trong một phiên cụ thể **u**

**Định nghĩa (Độ sâu phiên)** Tôi xác định độ sâu của phiên u là maxdepth (Pi, u) với Pi ∈ u. Tôi là quan tâm đến biến này vì phân phối của nó có liên quan theo quan điểm của các công cụ tìm kiếm. Để lướt ngẫu nhiên, Tôi có thể lập mô hình từng trang trong Pages như một trạng thái trong một hệ thống, và mỗi siêu kết nối trong Links như một quá trình chuyển đổi có thể xảy ra. Tôi đề xuất sử dụng một mô hình có liên quan có thể thu gọn nhiều trang ở cùng cấp độ như một nút đơn.

Ưu điểm của việc lập mô hình biểu đồ trang Web dưới dạng danh sách tuần tự vì chúng ta không cần lập mô hình chính xác người dùng đang truy cập trang nào bởi vì tôi không cần thông tin này vì mối quan tâm chính của tôi là độ sâu của người dùng bên trong một trang Web. Ngoài ra, các trang Web khác nhau có mức độ kết nối khác nhau, do đó, để xem xét toàn bộ trang Web, chúng ta cần mô hình hóa cả số lượng liên kết ngoài của mỗi trang và sự phân bố chồng chéo của các liên kết ngoài giữa các trang.

Ở mỗi bước truy cập, người tru cập có thể thực hiện một trong các hành động sau: đi tới cấp tiếp theo ( hành động đi tiếp), trở lại cấp phía trước ( hành đồng trờ lại), đứng yên tại nơi cùng cấp ( hành động ở lại), đi đến một bcấp độ trước đó khác nhau (hành động trước), đi đến một cấp độ khác sâu hơn, đi đến trang bắt đầu, nhảy ra ngoài trang web.

Đối với bước nhảy hành động, chúng tôi thêm một nút EXIT bổ sung để báo hiệu kết thúc phiên người dùng (đóng trình duyệt, hoặc đến một trang Web khác)

Vì nút EXIT này có một liên kết ra ngoài, nó không ảnh hưởng đến kết quả cho các nút khác nếu chúng ta loại bỏ nút EXIT và thay đổi điều này bằng cách chuyển tiếp đến mức bắt đầu L0. Another way to understand it is that as this process has no memory, quay lại trang bắt đầu hoặc bắt đầu một phiên mới là tương đương nhau, vì vậy các hành động nhảy và bắt đầu không thể phân biệt được về phân phối xác suất kết quả cho các nút khác

### 7.3 Mô hình A: quay lại từng cấp một

Trong mô hình này, với xác suất q, người dùng sẽ tiến sâu hơn và với xác suất 1 - q, người dùng sẽ trở lại một cấp độ.

Xác suất chuyển đổi được đưa ra bởi:

* Pr(next|) = q
* Pr(back| = 1−q for ≥ 1
* Pr(stay|) = 1−q for = 0
* Pr(start, jump|) = 0
* Pr(prev|) = Pr(f wd|) = 0

Trạng thái ổn định x được đặc trưng bởi:

xi = qxi−1 + (1−q)xi+1 (∀i ≥ 1)

x0 = (1−q)x0 + (1−q)x1

Giải pháp cho sự tái diễn này là:

Nếu q ≥ 1/2 thì nghiệm là xi = 0 và x∞ = 1, vì vậy chúng ta có một trạng thái hấp thụ tiệm cận. Trong khuôn khổ của tôi bđiều này có nghĩa là không có ranh giới độ sâu nào có thể đảm bảo một tỷ lệ trang nhất định được người dùng truy cập. Khi nào q <1/2 và chúng tôi áp đặt ràng buộc chuẩn hóa, ∑i≥0 xi = 1, Tôi có phân phối hình học:

Xác suất tích lũy của các mức 0 ... k là:

Tôi cũng tính toán thời lượng phiên, nếu chúng tôi cho rằng phiên kết thúc khi người dùng trở về mức 0, vì các hành động bắt đầu và nhảy tương đương nhau.

### 7.4 Dữ liệu từ các phiên người dùng trong các trang Web

Tôi đã lấy được nhật ký truy cập có địa chỉ IP ẩn danh từ các trang Web này và xử lý chúng để có được phiên người dùng:

* Sắp xếp nhật ký theo địa chỉ IP của máy khách, sau đó theo dấu thời gian truy cập.
* Chỉ xem xét các yêu cầu GET cho các trang hoặc tài liệu HTML tĩnh và động như Word, PDF
* Hãy xem xét rằng một phiên sẽ hết hạn sau 30 phút không hoạt động, vì điều này thường xảy ra trong phần mềm phân tích tệp nhật ký và dựa trên dữ liệu thực nghiệm
* Hãy xem xét rằng một phiên sẽ hết hạn nếu User-Agent thay đổi, như một cách khắc phục vấn đề nhiều máy khách có thể đứng sau cùng một địa chỉ IP.
* Xem xét nhiều lần truy cập liên tiếp vào cùng một trang (tải lại trang) dưới dạng một lần xem trang
* Xem xét nhiều lần truy cập liên kết vào cùng một trang (tải lại trang) dưới dạng một lần xem trang
* Bỏ qua các lần truy cập vào các ứng dụng Web như e-mail hoặc hệ thống quản lý nội dung, vì chúng không phản hồi đối với logic của việc duyệt trang, cũng như các trình thu thập dữ liệu Web thường không thể truy cập được.
* Mở rộng phiên với các trang bị thiếu (ví dụ: nếu người dùng nhấp vào "quay lại" trong trình duyệt của họ, sau đó làm theo liên kết). Thông tin này được lấy từ trường Liên kết giới thiệu và là một cách để khắc phục một phần vấn đề về bộ nhớ đệm. Lưu ý rằng, vì các lượt truy cập lại không phải lúc nào cũng được ghi lại do bộ nhớ đệm dữ liệu từ các tệp nhật ký đánh giá quá cao độ sâu mà người dùng đã sử dụng hầu hết thời gian, vì vậy số lượt truy cập của người dùng có thể thực sự thậm chí ít sâu hơn.

Ngoài ra, việc kiểm tra dữ liệu theo cách thủ công dẫn đến những suy đoán sau để loại bỏ các tác nhân tự động:

* Xác định rô bốt bằng cách chúng truy cập vào file robots.txt
* Xác định rô bốt bằng cách đã biết trường User-Agent
* Bỏ qua các lần truy cập độc hại tìm kiếm lỗ hổng bảo mật, thường là một chuỗi các yêu cầu tìm kiếm đối với lỗi tràn bộ đệm hoặc các lỗi phần mềm khác. Những yêu cầu này thường được thực hiện bởi các đại lý tự động như Nessus

Đặc điểm chung của phiên người dùng. Các đặc điểm của mẫu cũng như kết quả của việc lắp các mô hình B và C. Tên của các trang Web không được công khai vì một số tệp nhật ký đặc biệt là của các tổ chức thương mại, được thu thập trong điều kiện chỉ công bố các kết quả thống kê.

## 8. Đề xuất cho Hợp tác Máy chủ Web

Khi số lượng các trang web có sẵn công khai tăng lên, vấn đề giữ cho chỉ mục của công cụ tìm kiếm được cập nhật với các thay đổi ngày càng trở nên khó khăn hơn và việc tìm thấy một số trang đã cũ trong các công cụ tìm kiếm hiện tại là điều thường thấy. Điều này khiến mọi thứ trở nên khó khăn hơn đối với người dùng tìm kiếm thông tin và ảnh hưởng đến hình ảnh của công cụ tìm kiếm, nhưng điều này cũng có chi phí cho các trang Web bị xuyên tạc, nếu chúng ta xem xét toàn bộ trải nghiệm người dùng.

Nếu người dùng tìm kiếm một từ khóa trên công cụ tìm kiếm và sau đó chọn một trang từ kết quả tìm kiếm không còn tồn tại hoặc chứa tài liệu hiện không liên quan đến nhu cầu thông tin của người dùng, người dùng sẽ bực bội với cả công cụ tìm kiếm và trang Web vì không tìm thấy thông tin này. Ngoài ra còn có chi phí cơ hội liên quan đến những khách truy cập này: có thể thông tin họ muốn đã được chuyển đến một trang khác trong cùng một trang web và công cụ tìm kiếm không biết về sự thay đổi. Trong trường hợp này, nó sẽ tốt hơn cho Trang web để thông báo cho công cụ tìm kiếm về bản cập nhật.

Trình thu thập thông tin web có thể sử dụng một lượng tài nguyên mạng và bộ xử lý quan trọng từ việc phục vụ Web. Trình thu thập thông tin web có xu hướng truy cập nhiều nhiều trang hơn con người, and they request them very fast, normally with 10 to 30 seconds between visits. Nhiều yêu cầu là tài nguyên chưa được sửa đổi và có thể tránh được trong một số lược đồ nhất định nếu máy chủ thông báo cho trình thu thập thông tin về việc tài nguyên không được sửa đổi kể từ lần truy cập cuối cùng của nó.

Do đó, quản trị viên trang Web có các khuyến khích để cải thiện việc trình bày trang Web của mình trong chỉ mục của công cụ tìm kiếm và để ngăn chặn các lượt truy cập không cần thiết từ trình thu thập thông tin. Cơ chế hoàn thành đây là cái mà tôi gọi là sơ đồ hợp tác giữa máy chủ Web và trình thu thập thông tin web.

### 8.1 Đề án hợp tác giữa các máy chủ

Giao dịch HTTP tiêu chuẩn tuân theo mô hình yêu cầu-phản hồi: khách hàng yêu cầu một trang từ Web máy chủ và máy chủ Web phản hồi với trang. Giao dịch này liên quan đến siêu dữ liệu (thông tin về trang) được tải xuống cùng với nội dung trang thực tế.

Các phương án hợp tác mà tôi nghiên cứu trong luận văn này có thể được chia thành hai nhóm: thăm dò ý kiến và gián đoạn.

**Lược đồ thăm dò**: trình thu thập thông tin Web yêu cầu dữ liệu định kỳ từ máy chủ Web dựa trên chính sách lập lịch trình. Những yêu cầu này kiểm tra xem một trang đã được thay đổi hay chưa, sau đó tải xuống trang.

**Lược đồ ngắt**: trong trường hợp này, máy chủ Web bắt đầu giao dịch với công cụ tìm kiếm khi có sự kiện. Những sự kiện này có thể xảy ra khi một hoặc nhiều trang được cập nhật, dựa trên chính sách máy chủ. Các công cụ tìm kiếm phải đăng ký các máy chủ mà họ muốn nhận sự kiện. Điều này tương tự như mối quan hệ giữa bộ xử lý chính và thiết bị phần cứng (mạngthẻ, máy quét, v.v.) trong một máy tính hiện đại.

Lưu ý rằng cuộc bỏ phiếu trở nên tương đương với sự gián đoạn khi thời gian bỏ phiếu có xu hướng bằng không; nhưng cách sử dụng tài nguyên ở cả hai đầu đường bỏ phiếu tăng cùng một lúc. Trong luận án này, tôi nghiên cứu một số phương án hợp tác.

Trước khi tôi đi vào chi tiết của từng chương trình, có một số vấn đề tôi phải đề cập đến gần như là độc lập với chương trình được sử dụng.

**Nén** có thể được sử dụng để giảm chi phí truyền tải với chi phí xử lý nhiều hơn cấp nguồn ở phía máy chủ. Trên trang Web hiện tại, tính năng nén được sử dụng để truyền hầu hết các hình ảnh –vì chúng thường được lưu trữ ở định dạng nén - nhưng thông thường nó không được áp dụng cho các trang Web văn bản. Giao thức HTTP/1.0 xem xét các yêu cầu của các cơ quan nén bằng cách sử dụng Accept-encoding header, vì vậy việc nén các trang Web có thể được sử dụng nhưng nó thường được để cho giao tiếp giữa ISP và người dùng cuối. Nén có thể được sử dụng với các trang Web hoàn chỉnh, các gói trang Web và tài nguyên của chúng hoặc trang web khác.

**Các vấn đề về quyền riêng tư** phát sinh khi trình thu thập thông tin có quyền truy cập vào thông tin trong máy chủ mà không phải là công cộng. Điều này nghe có vẻ lạ, nhưng trong thực tế, khi sử dụng trình thu thập thông tin Web, bạn có thể tải xuống các tệp được liên kết do nhầm lẫn hoặc các thư mục riêng tư cho phép một danh sách ảo. Nhiều quản trị viên trang web nhầm tưởng rằng do không xuất bản URL của một trang Web, họ có thể giữ trang đó ở chế độ riêng tư. Đây là một thực tế phổ biến, vì vậy hầu hết các quản trị viên trang web rất miễn cưỡng cung cấp quyền truy cập cho khách hàng để liệt kê nội dung của các thư mục. Lưu ý rằng nếu người dùng theo dõi liên kết bên ngoài từ một trong những trang “riêng tư” này, trình duyệt của họ sẽ thông báo cho người giới thiệu đến trang Web tiếp theo, và liên kết giới thiệu này có thể được ghi lại và kiểm tra, vì vậy việc cố gắng giữ các URL không xác định ở chế độ riêng tư là vô ích.

**Khả năng cập nhật chỉ mục** rất giảm trong các công cụ tìm kiếm Web quy mô toàn cầu: các hạn chế về điều kiện không gian đĩa là giới hạn quan trọng nhất, vì vậy không phải lúc nào cũng có thể lưu trữ một bản sao hoàn chỉnh của các trang đã tải xuống. Điều này có thể dẫn đến một số khó khăn; ví dụ: trên một chỉ mục đảo ngược tiêu chuẩn, loại bỏ một trang hoặc cập nhật một đoạn văn mà không có văn bản hoàn chỉnh có thể là không thể hoặc rất tốn thời gian. Ngoài ra, trong nhiều trường hợp, cập nhật hàng loạt trang được ưu tiên hơn cập nhật các trang đơn lẻ để giảm tổng thể chi phí xử lý.

**Công cụ tìm kiếm “spamming”** xảy ra bất cứ khi nào các quản trị viên trang Web cố gắng đạt được mức cao không đáng có xếp hạng trong công cụ tìm kiếm. Ví dụ, dữ liệu do máy chủ Web cung cấp không thể được tin cậy hoàn toàn về mặt tần suất cập nhật trang tự khẳng định hoặc tầm quan trọng của trang địa phương. Trình thu thập dữ liệu web được hầu hết mọi người sử dụng công cụ tìm kiếm chỉ quan tâm đến một số trang của trang Web và quyết định trang nào phải được thêm vào chỉ mục nên được để lại cho công cụ tìm kiếm. Trong các kế hoạch thông báo, mức độ tin cậy có thể được thiết lập, ví dụ: nếu một trang Web gửi thông báo cập nhật, nhưng khi các trang được trình thu thập thông tin Web kiểm tra chúng không thay đổi, thì công cụ tìm kiếm có thể bỏ qua các thông báo khác từ trang Web đó trong một khoảng thời gian của thời gian.

**Cấu trúc HTML** được sử dụng trong một trang Web ảnh hưởng đến khả năng hiển thị các trang của nó bởi công cụ tìm kiếm trình thu thập thông tin. Thông tin chỉ có thể truy cập thông qua các biểu mẫu nói chung là khó thu thập cho trình thu thập dữ liệu Web; đây được gọi là “Web ẩn”. Các trang web có thể thu hút nhiều khách truy cập hơn nếu chúng cung cấp giao diện thân thiện với trình thu thập thông tin để dữ liệu này được các công cụ tìm kiếm lập chỉ mục

### 8.2 Hợp tác dựa trên lược đồ thăm dò

Trong tất cả các trường hợp này, trình thu thập thông tin Web truy vấn máy chủ Web với tính định kỳ nhất định;

**Thông tin meta-data được cập nhật.** Tệp chứa dữ liệu sửa đổi lần cuối (và có thể là kích thước tệp và đường dẫn) được phục vụ. Tệp này có thể chứa mô tả của nhiều trang trên trang Web. Trong trường hợp các tệp đơn lẻ, giao thức HTTP cung cấp HEAD yêu cầu được phản hồi kèm với meta-data liên quan về yêu cầu của đối tượng. Rất nhiều HEAD yêu cầu có thể được tạo ra, nhưng ddeièu này không hiểu quả bằng việc cung cấp một tệp ngắn gọn.

Cuối cùng, tiêu đề HTTP hết hạn: tiêu đề trình bay thông báo tới trình thu thập thông tin thay đổi tiếp theo trong một trang web, nhưng điều này liên quan đến dự đoán và do đó không được sử dụng rộng rãi.

**Hệ thống có sự khác biệt về nội dung.** Máy chủ Web cung cấp một loạt sự khác biệt giữa phiên bản cơ sở và phiên bản mới hơn. Trong trường hợp đơn giản nhất, sự khác biệt chỉ là giữa phiên bản cuối cùng và phiên bản hiện tại. Ví dụ: với một yêu cầu HTTP có sử dụng mã hoạt nội dung Content-encoding (thuộc tính của HTTP response). Tuy nhiên một nhược điểm là các máy chủ phải giữ lại nhiều phiên bản của trang web của họ và nó chỉ có thể sử dụng cho các trang web đã được truy cập. Một nhược điểm khác là số lượt truy cập chiếm một phần nhỏ trong tổng số lượt tải xuống, vì vậy điều này không thể được sử dụng cho tất cả các trang.

**Chỉ phân phát các trang nếu được sửa đổi.** Trình thu thập thông tin web có thể tránh tải xuống tệp nếu tệp chưa được sửa đổi.

Ví dụ trong HTTP/1.0 điều này thự hiện bằng cách sử dụng ngày mà trình thu thập cung cấp thông tin và trường If-Modified-Since trong header, tuy nhiên những tiêu ddề này chỉ được sử dụng bởi một số ít trình thu thập thông tin, mặc dù chúng được hỗ trợ hầu hết tại các máy chủ web

**Cung cấp nhiều trang theo một yêu cầu.** Có thể tránh được chi phí phát sinh từ nhiều kết nối TCP bằng cách yêu cầu một loạt các trang trong cùng một kết nối.

**Cung cấp toàn bộ trang web trong một tệp lớn.** Điều này chỉ phù hợp nếu các thay đổi trên Web xảy ra trong nhiều trang hoặc nếu trang web bao gồm nhiều tệp nhỏ.

**Lọc giao diện.** Đây là một phương pháp tiêu chuẩn để trả lời các truy vấn từ trình thu thập thông tin. Truy vấn điển hình trình thu thập thông tin web có thể yêu cầu là “cung cấp cho tôi tất cả các trang đã thay đổi kể từ ngày này”. Một mạnh mẽ hơn giao diện lọc cũng có thể bao gồm các yêu cầu về sự khác biệt hoặc truy vấn về các đặc điểm khác như kích thước trang hoặc tầm quan trọng của địa phương.

## 8.3 Hợp tác dựa trên lược đồ ngắt

Trong tất cả các trường hợp này, máy chủ Web gửi dữ liệu đến máy chủ Web bất cứ khi nào có bản cập nhật (trang web thay đổi, trang web bị xóa hay có một trang web mới). Trình thu thập thông tin web phải đăng ký với máy chủ Web để bắt đầu nhận thông báo, và khi nhận được, họ có thể chọn xử lý, xếp hàng hoặc bỏ qua.

**Gửi meta-data khi cập nhật.** Thông báo chỉ bao gồm siêu dữ liệu về bản cập nhật, tối thiểu, URL của tài nguyên và dấu thời gian của sự kiện sẽ được yêu cầu.

**Gửi những nội dung khác nhác nhau.** Bất cứ khi nào một sự kiện xảy ra, máy chủ Web sẽ gửi một tệp chứa sự khác biệt giữa phiên bản mới nhất và phiên bản hiện tại trên. Điều này khai thác thực tế là hầu hết các thay đổi trang là nhỏ.

**Gửi trang thay đổi:** Máy chủ Web gửi toàn bộ văn bản của mỗi trang được cập nhật hoặc trang mới khi sửa đổi được thực hiện.

**Gửi hang loạt các trang thay đổi:** Máy chủ Web gửi một loạt các trang đã sửa đổi theo một số lịch trình. Điều này có thể hữu ích nếu các bản cập nhật thường xuyên và liên quan đến một số trang; ví dụ, trong trang Web hàng ngày hoặc báo hàng tuần

**Gửi toàn bộ website.** Máy chủ Web gửi toàn bộ trang web. Điều này rất hữu ích, chẳng hạn, để tải lên toàn bộ trang web khi trang web được cung cấp công khai lần đầu tiên hoặc nếu có một sửa đổi lớn liên quan đến hầu hết các trang, như một phần mở rộng của lược đồ trước đó.

**Điều khiển từ xa.** Máy chủ Web thực thi phần mềm do công cụ tìm kiếm cung cấp; phần mềm này bao gồm hướng dẫn để xác định các trang quan trọng và phát hiện các thay đổi trong các trang có liên quan đến tìm kiếm. Các trang quan trọng có thể được xác định dựa trên kết nối cục bộ, thông tin dạng văn bản hoặc phân tích tệp nhật ký. Các thay đổi có thể được phát hiện bằng cách sử dụng một thuật toán tùy chỉnh khác nhau tùy thuộc vào đặc tính của công cụ tìm kiếm. Khi có thay đổi, agent sẽ gửi lại một số dữ liệu cho bộ máy tìm kiếm. Đây có thể là siêu dữ liệu, các trang hoàn chỉnh hoặc các điểm khác biệt.

## 9. Triển khai Crawler

### 9.1. Môi trường lập trình và các phụ thuộc

Ngôn ngữ sử dụng là javascript sử dụng nodejs để khởi chạy và build code. Thư viện sử dụng **puppeteer** dùng để tải website và thao tác website.

Các trang web tập trung gồm các trang thuộc lĩnh vực báo chí có tên miền .vn ví dụ: dantri.com.vn, 24h.com.vn

### 9.2 Về chương trình

Trong một phiên làm việc sẽ có bốn chương trình chính chạy trong một vòng đời của việc thu thập thông tin: Quản lý, Tìm kiếm, Phân tích, tổng kết

#### 9.2.1 Quản lý

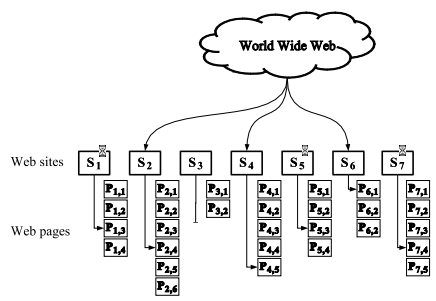
Chương trình sẽ tự động sinh ra một danh sách các URL mà cần thu thập thông tin. Dựa vào các yếu tố khác nhau:

1. Lọc ra các trang đã được tải xuống quá gần đây. Trong các tệp cấu hình sẽ chứa tần suất trang web cần thu thập theo tuần hoặc theo ngày hay theo tháng. Dựa thời yếu tố thời gian trình quản lý sẽ lọc các thông tin không cần thiết.
2. Ước lượng giá trị của trang web: Chương trình quản lý tính toán giá trị của tất cả các trang Web các trang trong bộ sưu tập theo chức năng xếp hạng sắp xếp.
3. Ước tính độ mới của các trang web: Chương trình khởi tạo và gán cho mỗi trang web một giá trị thay đổi là (p=1) xong đó sẽ tổng hợp dựa vào các kết quả để tính toán.

#### 9.2.2 Tìm kiếm

Chương trình sẽ lựa chọn từ trong các danh sách mà trình quản lý yêu cầu và tiến hành phân tích tìm kiếm các thông tin cần thiết.

Các thông tin đa phần kiểm tra xem website có thay đổi gì mới hay có điều gì mới hay không, liệu website có trang nào bị xóa. Chương trình này thường sẽ thao thác phân cấp với các mạng lưới của một website như Hình 8.1



Hình 8.1: Mô hình một trang web cụ thể sẽ có phân cấp và mở rộng

Chương trình tìm kiếm đôi lúc sẽ sử dụng đa luồng để tìm kiếm các trang của web một cách nhanh nhất.

Sau khi tìm kiếm xong sẽ thông báo về cho quản lý và quản lý sẽ có nhiệm vụ lọc ra những trang web nào thay đổi và cần phải cập nhật

#### 9.2.3 Phân tích

Chương trình phân tích sẽ nhận các bản rõ về từng trang của web đã được chọn lọc và tải xuống từ trước. Và cách cài đặt hiện tại tôi chỉ thực hiện bọc tách các đoạn text cần thiết cho văn bản.

Việc phân tích trang web có chứa mã code HTML sẽ thực hiện bằng các cũ pháp của document để thực hiện lấy thông tin giữa các thẻ

Trong thời gian phân tích Chương trình sẽ phát hiện và sẽ thêm vào các trạng thái khác nhau của phần “tổng kết” các thông tin như là tại URLS xuất hiện thông tin bị trùng lặp.

Sau khi hoàn thành việc phân tích lương dữ liệu còn lại đã được chuẩn hóa và đựa vào để sử dụng lưu trữ và bóc tách để phục vụ các chức năng khác nhau.

#### 9.2.4 Chương trình Tổng kết

Chương trình sẽ nhận thông báo từ quá trình phân tích để đưa tới kết luận với từng URLS

**Accept** quá trình bóc tách phân tích dữ liệu hoàn toàn phù hợp với trang web này và tên miền cùng loại. sửa đổi thông tin cho phép tiếp tụp cập nhật các url đó và đồng thời có thể sử dụng tìm kiếm trên URL đó

**Reject** quá trình bóc tách phân tích dữ liệu xảy ra lỗi nên việc cập nhật url hay tải về sẽ bị ngưng lại. đồng thời cũng không thể tiến hành tìm kiếm trên url đó

### 9.3 Vấn đề an toàn

Các vấn đề an toàn luôn luôn tồn tại ngay cả khi hệ thống crawling chỉ thực hiện trong mạng nội bộ. Tuy nhiên lượng dữ liệu thu thập được lại public ra ngoài.

**Ảnh hưởng tính sẵn sàng:** Do hệ thống thu thập là phụ thuộc nhiều vào một trang web trên internet do đó việc trang web kia bị sập hay bị hỏng là điều không thể dự đoán trước. Hay chỉ đơn giản là thay đổi cấu trúc HTML cũng gây ảnh hưởng tới hệ thống thu thập. Cho nên cần tách biệt hệ thống thu thập và cần xử lý tốt việc thông báo lỗi và không ngưng hệ thống.

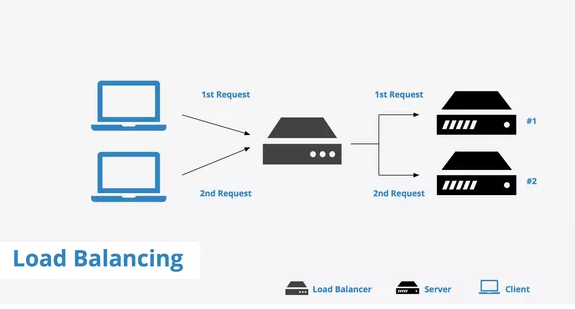
**Ảnh hưởng về các lỗi tấn công XSS.** Tuy rằng hệ thống hoàn toàn ko cung cấp giao diện cho người dùng và cũng không cho người dùng sửa đổi dữ liệu về của hệ thống. Tuy nhiên điều này không có nghĩa là các trang web mà hệ thông tiến hành thu thập thông tin không bị tấn công. Do đó mỗi lần thu thập thông tin đều phải lọc bỏ các kí tự đặc biệt đề phòng bị ảnh hưởng các cuộc tấn công XSS từ các trang web khác.

**Ảnh hưởng tới việc chiếm lưu lương băng thông.**  Hệ thống thu thập cần rất nhiều băng thông để phục vụ cho mục đích tải các trang mạng mới về và tiến hành cập nhật các trạng thái. Việc này diễn ra thường xuyên và liên tục cho nên sẽ chiếm khá nhiều băng thông và lưu lượng mạng. việc đó có khả năng khiến cho trang chính không sử dụng đc mạng để truyền tải dữ liệu cho user.

# CHƯƠNG 5: TRIỂN KHAI CÁC BIÊN PHÁP AN TOÀN

## 1. Tích hợp Load balancing

Khái niệm về Load Balancing (Cân bằng tải): là tính năng giúp máy chủ ảo hoạt động đồng bộ và hiệu quả hơn thông qua việc phân phối đồng đều tài nguyên. Việc cân bằng tải là một trong những tính năng quan trọng với những nhà phát triển, lập trình mạng. Khi người dùng truy cập vào một trang web có nhiều vấn đễ có thể xảy ra như web server không kịp xử lý, tải chậm, hay không kết nối được tới máy chủ…. Và tác dụng của Load Balancing xử lý các sự cố này. Cơ chế hoạt động của Load Balancing là khi người dùng truy cập tới Load Balancer thì Loađ Balancer sẽ chuyến tới một máy chủ khác để thực hiện tác vụ. Dù máy chủ chính bị lỗi hoặc nghẽn mạng thì tất cả các yêu cầu người dùng đều được giải quyết.



Hình 3.5: Cơ chế hoạt động của Load Balancing

* Công cụ sử dụng **Nginx** và cơ chế **Round Robin**

Khái niêm về Nginx: là một phần mềm web server mã nguồn mở nổi tiếng. Ban đầu nó dùng để phục vụ web HTTP. Tuy nhiên, ngày nay nó cũng được dùng làm reverse proxy, HTTP load balancer và email proxy như IMAP, POP3, và SMTP.

Khái niệm về round robin: là thuật toán lựa chọn các máy chủ theo trình tự. Theo đó, Load Balancer sẽ bắt đầu đi từ máy chủ đầu tiên trong danh sách của nó và ứng với yêu cầu đầu tiên. Tiếp đó, nó sẽ di chuyển dần xuống trong danh sách theo thứ tự và bắt đầu lại ở đầu trang khi đến máy chủ cuối cùng. Round robin thường được cài đặt cùng với các phương pháp duy trì session như sử dụng cookie.

* Triển khai cấu hình:

1. Cấu hình nginx trong file config:   
   

Hình 3.6: Triển khai cấu hình routing cho website

Cấu hình server là yêu cầu nginx lắng nghe cổng 80 và redirect những domain là localhost về backend để xử lý các request được nhận.

Kết quả: Với mỗi lần người dùng truy cập tới hệ thống. Load Balancer (Nginx) sẽ điều hướng vào server localhost:8081 với request đầu tiền, server localhost:8083 với request thứ 2 và lặp lại. Khi triển khai hệ thống sẽ có lượng chịu tải được nhiều hơn và khi mở rộng hơn thì chúng ta cần triển khai source code lên một port khác và khai báo thêm vào trong danh sách của nginx

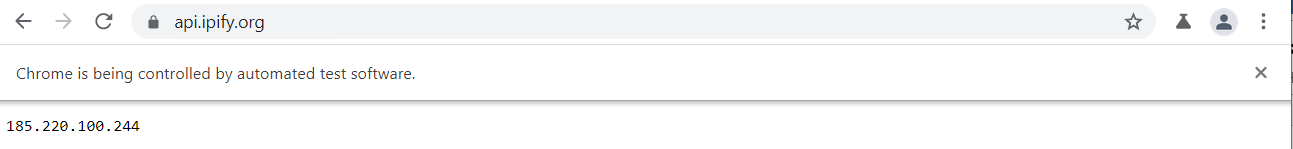
## 2. Tích hợp tor proxy vào để thực hiện crawling

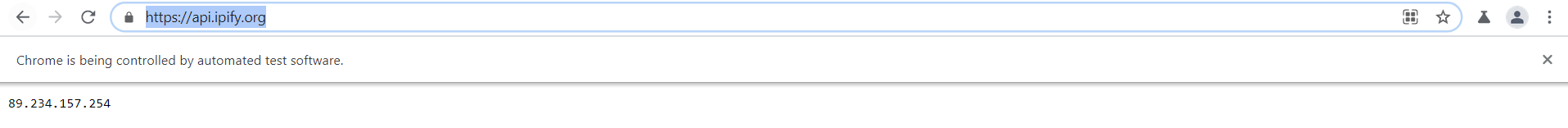
Proxy server hoạt động như một cổng nối giữa người dùng và internet. Đây là một server trung gian giữa người dùng cuối và trang web họ truy cập. Các máy chủ proxy cung cấp các chức năng, bảo mật và riêng tư khác nhau phụ thuộc vào chính sách.

Bài toán đăt ra: khi thực hiện crawling lương truy cập tới một website cụ thể rất liên tục khiến cho website nghĩ ngờ minh là đối tượng tấn công ddos cho nên cần thay đổi ip cho mỗi lần request tới website để crawling dữ liệu.

Từ đó ta lựa chọn Tor như một proxy để thông qua đó mỗi lần truy cập tới website để thực hiện việc crawling sẽ sử dụng một địa chỉ ip mới.

Khi truy cập lần thứ nhất có sử dụng tor proxy tới trang web <https://api.ipify.org/> trang web sẽ hiển thị publish ip của bạn hiển thị ip là 185.220.100.244



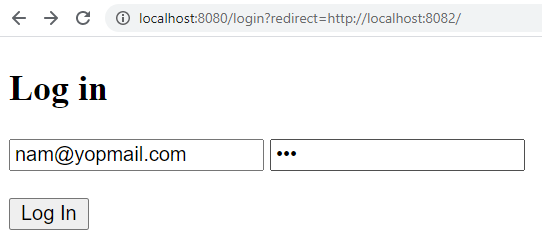
Khi truy cập lần thứ hai tới trang web <https://api.ipify.org/> trang web sẽ hiển thị publish ip của bạn hiển thị ip là 89.234.157.254  


Hình 3.7: Kết quả sử dụng proxy

Nhờ viêc thay đổi ip liên tục như vậy nên che giấu được thông tin bản thân thông qua các proxy. Dễ dàng vượt qua việc block ip của một trang web bất kì.

## 3. Thực hiện việc phân quyền cho hệ thống

- Hệ thống sẽ bao gồm 2 người dùng tương ứng với 2 quyền là ADMIN và USER. Nhằm bảo đảm an toàn thì người dùng có quyền là USER chỉ đươc phép truy cập vào trang web tin tức.



Hình 3.8: Người dùng đăng nhập thất bại vào trang quản trị

Với cách xử lý này sẽ hạn chế quyền của người dùng thao tác ngoài ý muốn vào những dữ liệu nhạy cảm. Có thể tách biệt hoàn toàn trang web của người dùng và trang web của người quản trị.

Và trang quản trị cũng không cần thiết chứa những chức năng của trang web người dùng từ đó có thể tập trung vào quản lý bài viết và phát triển chức năng dành riêng cho quản trị viên.

* Hệ thống xác sẽ thực hiện tạo cookie cho từng trang web và mỗi cookie sẽ kèm theo từng token riêng biệt. Sử dụng Javascript web token để tạo ra các token cho từng trang web. Đảm bảo việc một website bị tấn công đánh cắp cookie thì cookie đó cũng không thể sử dụng cho website khác trong cùng hệ thống

## 4. Ngăn chặn các tấn công khác

* Các tấn công sql injection: hệ thống sử dụng cơ sở dữ liệu mysql và mỗi câu truy vấn sẽ đều sử dụng prepared statements để đảm bảo các tham số truyền vào môt cách an toàn.
* Xss các lỗi về khi thu thập dữ liệu các kí tự đặc biệt sẽ chuyển đổi về kí tự hợp lệ trên html nhằm giảm thiểu ảnh hưởng tấn công Xss từ các website khác

# KẾT LUẬN

Ba chương của đồ án đã thể hiện được rằng những mục tiêu đặt ra khi thực hiện đồ án đều đã đạt được. Cụ thể:

Chương 1 đã nêu ra được thực trang vấn đề về thông tin hiện nay và tính cần thiết của đồ án.

Chương 2 nói tới viêc khảo sát và phân tích thiết kế hệ thống đối với một trang website tin tức. Các sơ đồ hoạt động của một trang website và chức năng của website

Chương 3 nói tới việc triển khai một website tin tức các mô hình cơ sở dữ liệu.

Chương 4 thực hiện phân tích các vấn đề liên quan tới thu thập dữ liệu vả các việc khó khăn hay thách thức trong quá trình thu thập dữ liệu. Kèm thêm vào đó là mô hình ứng dụng vào trong thực tế.

Chương 5 nêu ra thành quả cài đặt các chức năng và cách thức triển khai việc thu thập thông tin môt cách hiệu quả và không ảnh hưởng tới chương trình chính. Kèm theo đó là các cách thức để bảo đảm an toàn cho website thông tin.